MBA Business Foundations, Quantitative Methods: Session Five

Boris Babic,
Assistant Professor of Decision Sciences

INSEAD
The Business School for the World®

Today

R vs. Excel

- Steeper learning curve (whether it's worth it, depends on your objectives!)
- Fast
- Free
- Flexible
- Reproducible
- Extremely well supported (stackexchange, rbloggers, etc.)
- Big data friendly (eg. no bound on rows/columns)
- Visualization and graphics! (with ggplot2)
- Data science (with the Tidyverse)
- Machine learning (with Caret, randomForest, nnet, e1071)
- Bayesian modeling/Markov Chain Monte Carlo (with Bugs, Jags, and Stan)

… you learn as you go/as needed, shouldn't try to "learn R"!

Summary so far

- Into to functions (ordinary, inverse, quadratic...)
- Exponents and logs
- Derivatives
- Optimization
- Statistics and probability (last class: measures of location and dispersion)

Probability

- Tyche: Goddess of chance (daughter of Zeus).
- The ancient Greeks believed that when no other cause can be attributed to random events such as floods, droughts, frosts, then Tyche is responsible.
- Probability is the study of such random events or, more generally, of randomness.

Probability

- But what is probability?
- For our purposes, the probability of an event can be interpreted as the long-run frequency of the occurrence of that event.
- Always???
- What is the probability of a nuclear war in the next year? Of US dollar collapse?
- When Galileo first observed Saturn through a telescope, he saw something like this.

- Are those rings around the planet? Handles? Or is it three planets next to each other? Can you assign a probability?

Probability vs. statistics

- Revisit: What is the difference between probability and statistics?
- A probability question: A fair die will be tossed twice. What is the probability that it lands on six both times?
- A (descriptive) statistics question: A die was tossed twice and it landed on a five and a six. What is the mean die value?
- An (inferential) statistics question: A die was tossed twice and it landed on a five and a six. How confident are you that the die is fair?
- To answer this question, can you use descriptive statistics? Can you use probability? How? Is there a right answer? What is it?

Random variables

- In probability we know which values our variables X can take, and we know how probable those values are.
- Ex: if the variable represents the outcome of the toss of a fair die (i.e., which face landed up), what are the values? How likely are they?
- Variables are
  - Discrete, corresponding to natural or counting numbers.
  - Continuous, corresponding to real numbers.
- Classify the following:
  - Height or weight
  - Number of monthly lottery winners in California
  - Temperature tomorrow
  - A parent's number of children
  - The amount of money in your bank account

Knowing your sample space

- A certain couple has two children. At least one of them is a boy. What is the probability that both children are boys?
- Possibilities: BB, BG, GB, GG
- What can we rule out? GG
- What remains: BB, BG, GB
- Probability that both children are boys is $1/3$.

Probability distributions

- A probability distribution is a function f of the random variable X: $f(X)$
- This function can only take values between 0 and 1.
- Also, this function is additive: for two independent events, the probability of their sum is the sum of their probabilities.
- Ex: Pr (die lands on 4) + Pr (die lands on 6) = Pr (die lands on 4 or die lands on 6).
- In the discrete case, it tells us how probable it is that the random variable X will take a specific value x. We often denote the function f with $\Pr$.
- Ex: For a fair die, $f(3) = \Pr(X = 3) = 1/6$.
- In the continuous case, it tells us how likely it is that our variable will be contained within an interval $[a, b]$.
- Ex: For a person's weight, $\Pr(a < x < b) = k$.
- But what about $\Pr(x)$ in the continuous case (???).

Variations on exercise problem 4

- A fair die is rolled once.
- What is the probability that it lands on a number greater than 4?

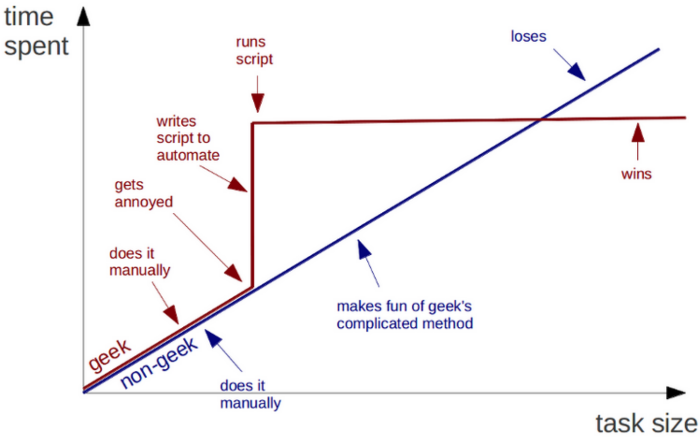
$$\Pr(X > 4) = \Pr(X = 5) + \Pr(X = 6) = 2/6 = 1/3$$

```
library(dice)
getEventProb(nrolls = 1,ndicePerRoll = 1,
nsidesPerDie = 6,eventList = list(5:6))
[1] 0.3333333
```

- A fair die is rolled 5 times. What is the probability of seeing exactly the pattern 6, 5, 4, 3, followed by a 2 or a 1?
- $\left(\frac{1}{6}\right)^5 + \left(\frac{1}{6}\right)^5 = 0.0002$

```
library(dice)
getEventProb(nrolls = 5,ndicePerRoll = 1,nsidesPerDie = 6,
eventList = list(6, 5, 4, 3, 1:2),orderMatters = TRUE)
[1] 0.0002572016
```

Doing things manually vs. automating them: geek vs. non-geek



Birthdays

- We have 30 students in this class. What is the probability that at least one pair of students share the same birthday?

- Algebraic solution:

  $\frac{30 \cdot 29}{2} = 435$ pairs of students

  $\frac{364}{365}$ probability that a single pair does not share a birthday

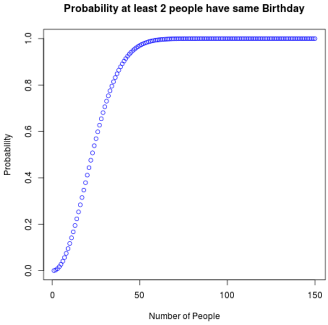
  $\left(\frac{364}{365}\right)^{435} = 0.30$ probability that no pair shares a birthday

  $1 - 0.30 = 0.70$ probability that at least one pair shares a birthday

Birthdays: R solution

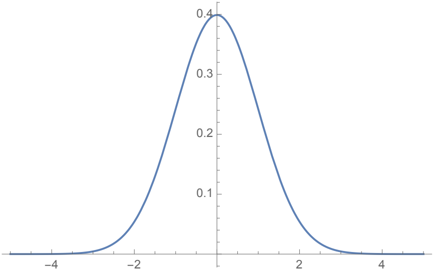
```
k = 30
p <- numeric(k)  # create numeric vector to store probabilities
for (i in 1:k)      {
            q <- 1 - (0:(i - 1))/365  # 1 - prob(no matches)
            p[i] <- 1 - prod(q)  }
prob <- p[k]
print(prob)
#BONUS:

plot(p, main="Probability at least 2 people have same Birthday",
xlab ="Number of People", ylab = "Probability", col="blue")
[1] 0.7063162
```



Normal distribution

- Most important probability distribution you will encounter (due, in part, to the central limit theorem).
- This distribution belongs to the exponential family of distributions, and it has two parameters, its average μ and standard deviation σ.
- Represented by the famous "bell curve": symmetric around its mean

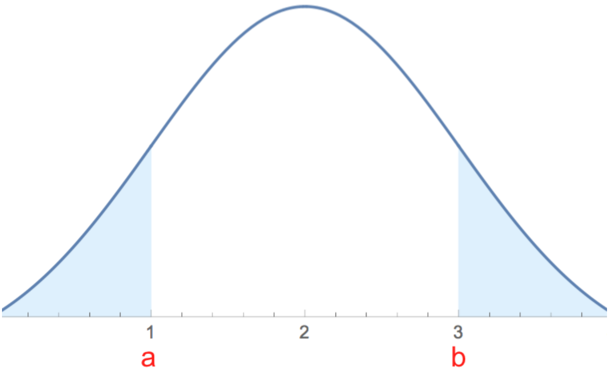


- Given by

$$f(x|\mu,\sigma^2) = (2\pi\sigma^2)^{-1/2} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

- Bonus: Why is π in there? Because $\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$!

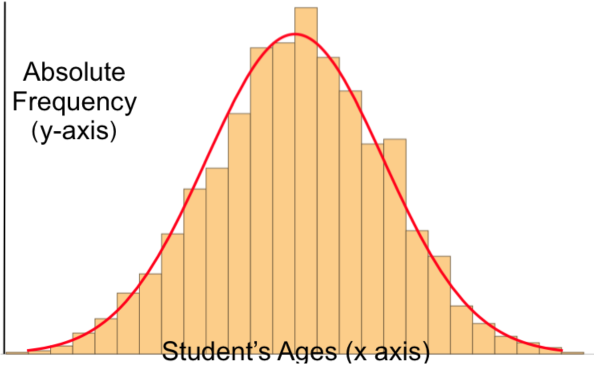
Normal distribution



The probability that our random variable is between a and b is given by the area under the curve between those two points:

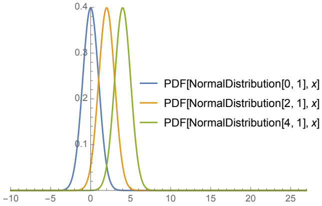
$$\Pr(a < x < b) = (2\pi\sigma^2)^{-1/2} \int_a^b e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$$

Normal distribution

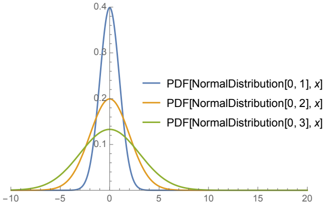


- Variables that have a normal distribution are ubiquitous in real life, provided we have enough data.
- Ex: Age of INSEAD students, height of INSEAD students.

Normal distribution



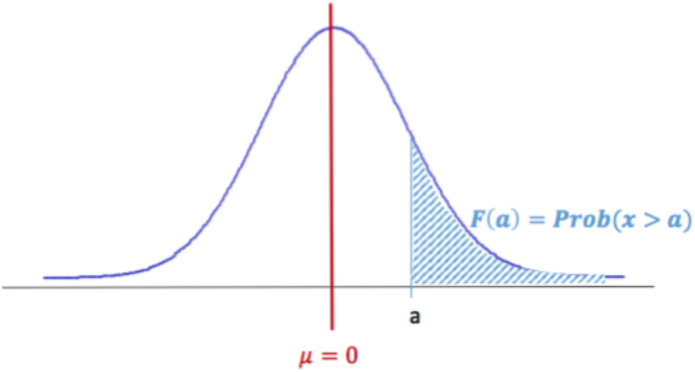
- As the mean changes, the location of the bell shifts
  To the left (for smaller means)
  To the right (for larger means)



- As the standard deviation changes, the bell becomes
  taller and thinner (for smaller standard deviations)
  shorter and thicker (for larger standard deviations)

Normal distribution

- So how to compute these areas under the curve (=probabilities)?
- The integral does not have a closed form.
- Rescale to a standard normal distribution and then use a table.
- Or, use computational approach, for example in R!



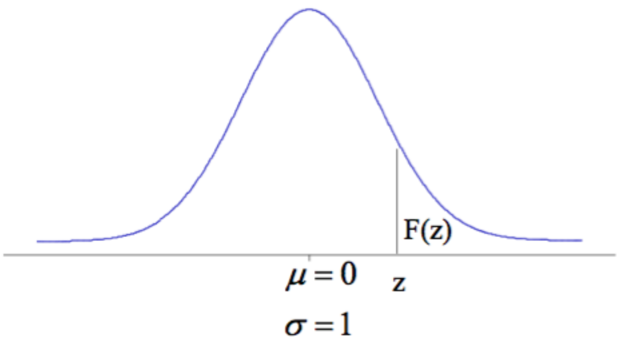
Transformations

- A standard normal distribution is a normal distribution with mean 0 and variance 1.
- The distribution function of a standard normal is given by

$$(2\pi)^{-1/2}e^{\frac{-x^2}{2}}$$

- To denote that a random variable X has a normal distribution we will use $X \sim \mathrm{N}(\mu, \sigma^2)$.
- if X follows a normal distribution with mean μ and standard deviation σ, $X \sim \mathrm{N}(\mu, \sigma^2)$, then $Z = \frac{x-\mu}{\sigma}$ follows a standard normal distribution, $Z \sim \mathrm{N}(0, 1)$.

Transformations

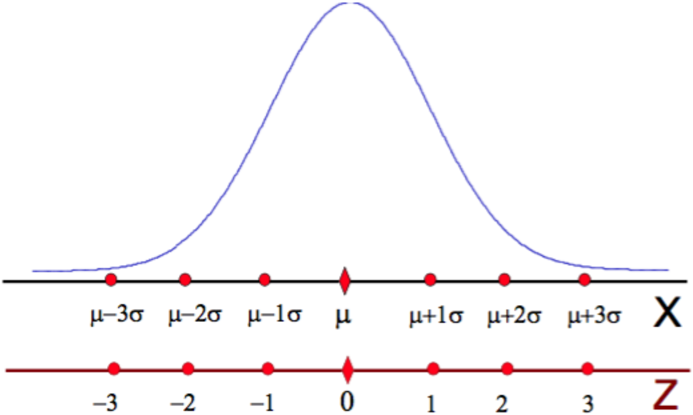


- Capital F denotes a cumulative distribution:

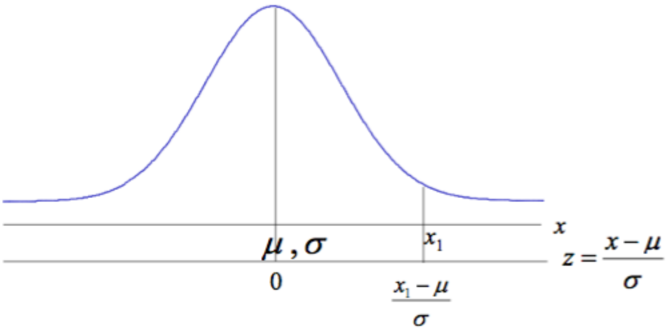
$$F(z) = Pr(Z \leq z) = \int_{-\infty}^{z} f(z)dz$$

- So $\Pr(a < x < b) = F(b) - F(a)$

Transformations



Transformations



Computing normal probabilities

- Ex: You want $\Pr(X > k)$ where $X \sim \mathrm{N}(\mu, \sigma)$.
- Step 1: Transform $X \to Z$

$$\begin{aligned}\Pr(X > k) &= \Pr\left(\frac{X-\mu}{\sigma} > \frac{k-\mu}{\sigma}\right)\\ &= \Pr\left(Z > \frac{k-\mu}{\sigma}\right)\end{aligned}$$

- Step 2: Look up the probability of $F(\frac{k-\mu}{\sigma})$ in R or a standard normal table.
- Step 3: Get the result:

$$\Pr(X > k) = F(\frac{k-\mu}{\sigma})$$

Examples

- Suppose that the test scores of a course exam at INSEAD are normally distributed with a mean of 72 and a standard deviation of 15.2. What is the probability that a randomly chosen student received above 84?

$$\Pr(X > 84) \to \Pr\left(\frac{X-\mu}{\sigma} > \frac{84-72}{15.2}\right) \to \Pr(Z > 0.789)$$

- pnorm(0.789, lower.tail=FALSE)
- 1-pnorm(0.789)
- OR, you can avoid the transformations altogether! pnorm(84, mean=72, sd=15.2, lower.tail=FALSE)
- Approximately 21%.
- We use lower.tail=FALSE in order to get the area from x to ∞.

The pnorm() function

- The pnorm function replaces the lookup table at the end of all statistics textbooks.
- pnorm returns the integral from $-\infty$ to k of the pdf of the normal distribution. That is, $F(k)$.
- If you do not set any further values, then k is a Z-score by default.
- However, you can specify the mean and variance as pnorm(2, mean $=$ 5, sd $=$ 3)

Practice

- The weekly salaries of the employees of a large corporation are assumed to be normally distributed with mean \$450 and standard deviation \$40.
- What is the probability that a randomly chosen employee earns more than \$500 per week?

Solution

- Find the Z score and look it up:

$$\Pr(X > 500) = \Pr\left(\frac{X-450}{40} > \frac{500-450}{40}\right) = \Pr\left(Z > \frac{5}{4}\right)$$

- pnorm(500, mean=450, sd=40, lower.tail=FALSE)
- approximately 10%.

Tips

- Probabilities correspond to areas
- Probabilities sum to 1: $\Pr(Z < k) = 1 - \Pr(Z > k)$
- Symmetry: $\Pr(Z < -k) = \Pr(Z > k)$
- For intervals, use substraction: $\Pr(a < Z < b) = \Pr(Z > a) - \Pr(Z > b)$

Practice: but how much money will I make???

- The average post INSEAD MBA starting salary is $170k$ with a standard deviation of $30k$.
- You want to know the probability of earning between 150 to $200k$ upon graduation.
- So you want to calculate the probability that a randomly chosen post-MBA INSEAD student starts at $150 - 200k$.
- Find the answer algebraically, then confirm in R.

Solution

Algebraic:

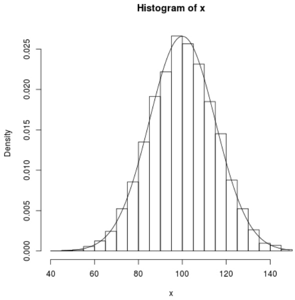
$$
\begin{aligned}
\Pr(150 < x < 200) &= \Pr(x > 150) - \Pr(x < 200) \\
&= \Pr\left(\frac{x-170}{30} > \frac{150-170}{30}\right) - \left(\frac{x-170}{30} > \frac{200-170}{30}\right) \\
&= \Pr\left(Z > \frac{150-170}{30}\right) - \left(Z > \frac{200-170}{30}\right) \\
&= \Pr\left(Z > -\frac{2}{3}\right) - \left(Z > 1\right) \\
&= .7454 - .1587 \\
&= .5867
\end{aligned}
$$

In R:

```
pnorm(150, mean=170, sd=30, lower.tail=FALSE) -
pnorm(200, mean=170, sd=30, lower.tail=FALSE)
```

The rnorm() function

```
x <- rnorm(10000, mean=100, sd=15)
hist(x, probability=TRUE)
xx <- seq(min(x), max(x), length=150)
lines(xx, dnorm(xx, mean=100, sd=15))
```

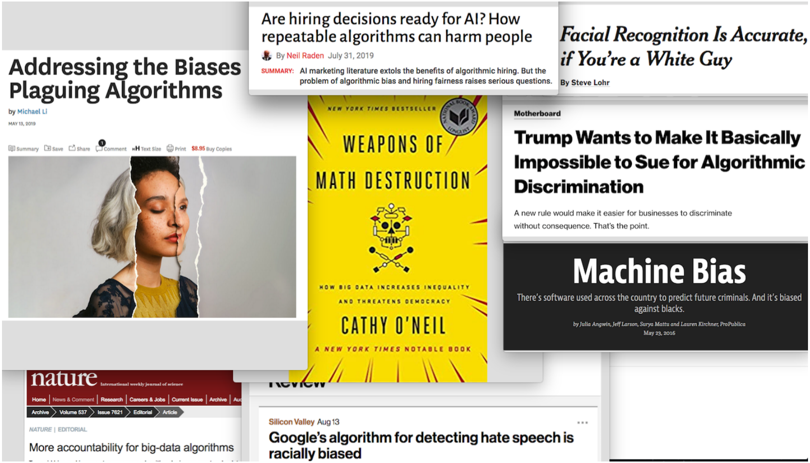


This generates 10000 random numbers from a specified normal distribution (first line), plots their histogram (second line), and graphs the distribution function of the same normal distribution (third and fourth lines).

Management decision making

- I will teach the MBA elective Management Decision Making (or MDM) in P4 (March/April)!
- You should all take it! Tell your friends too!
- Based on prototype I designed at Caltech called “Statistics, Ethics and Law”

Management decision making



Management decision making

Some topics we will look at:

- Human bias
  - Risk and rational choice
  - Statistical reasoning (p-hacking, small samples, overfitting)
  - Implicit bias, anchoring, loss aversion, ambiguity
  - Moral reasoning
- Machine bias
  - Algorithmic justice (ex: facial recognition systems)
  - Discrimination in lending, credit scoring, policing
  - Bias in advertising
  - Transparency in algorithmic decision-making (cf: adversarial attacks)
  - Privacy in data collection (for example in health care, on social media)
  - Ethics of self-driving cars and medical technology

Today

Basics	Functions Linear Inverse Two equations Quadratic
Exponents	Exponents Application: interest rates Exponential functions Logarithmic functions
Logarithms	Logarithmic functions Logarithmic and exponential equations Case: pricing Derivatives
Derivatives	Optimal decisions Case: production Statistics
Uncertainty	Probability & statistics Normal distribution

The Business School for the World®

Europe | Asia | Middle East