Calculating Probabilities with the Normal distribution

David Gerard

2017-09-18

Learning Objectives

- Standardizing Variables
- Normal probability calculations.
- Section 3.1 of DBC

Standardizing Variables

NBA Data

Player statistics for the 2016-2017 season of the NBA

- `player` The name of the player.
- `pts` The total points for the season
- `two_pp` Two point field goal percentage.
- `three_pp` Three point field goal percentage.
- Many others ...
- Here, I only kept players that attempted at least 20 two-point and 20 three-point field goals.

NBA Data

```
library(tidyverse)
nba <- read_csv("../../data/nba2016.csv") %>%
  filter(two_pa >= 20, three_pa >= 20) %>%
  select(player, pts, two_pp, three_pp)
glimpse(nba)
```

```
Observations: 337
Variables: 4
$ player   <chr> "Russell Westbrook", "James Harden", "...
$ pts      <int> 2558, 2356, 2199, 2099, 2061, 2024, 20...
$ two_pp   <dbl> 0.459, 0.530, 0.528, 0.524, 0.582, 0.4...
$ three_pp <dbl> 0.343, 0.347, 0.379, 0.299, 0.367, 0.3...
```

LeBron James

- LeBron James is the greatest player in the history of basketball (you will be tested on this).
- Is he better at three point field goals or two point field goals relative to other players?
- His three-point field goal percentage is 0.363 and his two-point field goal percentage is 0.611.
- Can we just say that he is a better two-point field goal shooter?

Not as easy as you think

- Can't just compare the numbers — three point field goals are much harder.
- I.e. the two statistics are in different units. We need a way to compare these observations without units.
- He *might* be better than most people at three point FG and worst than most people at two point FG, or vice versa.

Standardizing and z-scores

standardizing and z-scores

If x is an observation from a distribution that has mean μ and standard deviation σ, the standardized value of x is

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

A standardized value is often called a z-score.

The z-score is in units of standard deviations above the mean.

Mean and SD of two and three FG %

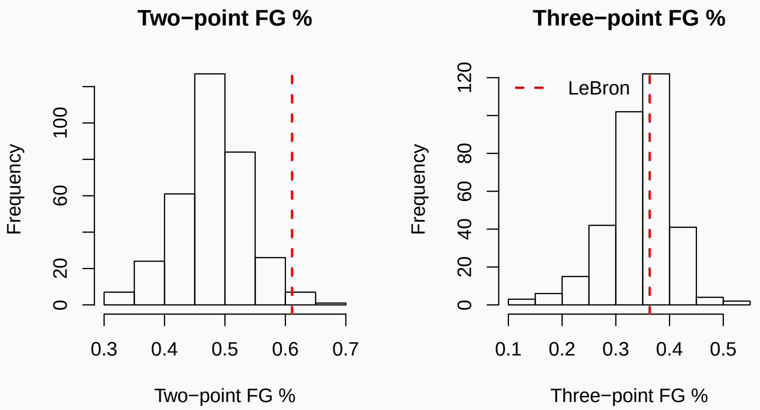
```
mu2    <- mean(nba$two_pp)
sigma2 <- sd(nba$two_pp)
mu3    <- mean(nba$three_pp)
sigma3 <- sd(nba$three_pp)
c(mu2, mu3)

[1] 0.4802 0.3431

c(sigma2, sigma3)

[1] 0.05779 0.05827
```

Three point field goals are harder!

LeBron's z-scores

- $z_2 = \frac{0.611 - 0.4802}{0.0578} = 2.2637.$
- $z_3 = \frac{0.363 - 0.3431}{0.0583} = 0.3414$
- The King (LeBron) is 2.26 SD's above the mean for two-point field goals but only 0.34 SD's above the mean for three-point field goals.
- Relative to everyone else, he is a lot better at two-point field goals.

Graphically



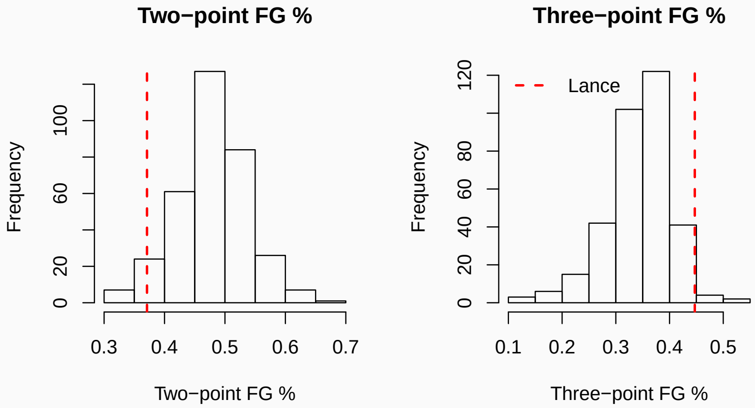
Another Example: Lance Thomas

- Lance Thomas (New York Knicks) has a two-point FT % of 0.371 and a three-point FG % of 0.447.
- Is he better at two-point field goals or three point field goals relative to his peers?

```
(0.371 - mean(nba$two_pp)) / sd(nba$two_pp)

[1] -1.889

(0.447 - mean(nba$three_pp)) / sd(nba$three_pp)

[1] 1.783
```

He is **way** better at three point field goals.

Graphically



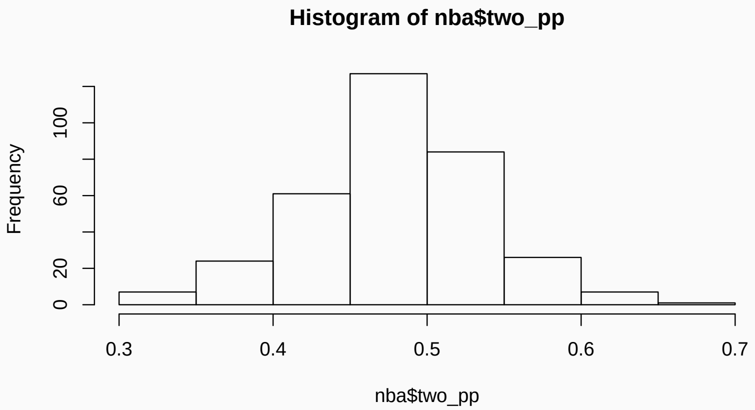
Other Examples

- Comparing the heights of two children of different ages ("which one is taller relative to their age?").
- Did you do better on the SAT or the ACT?
- How about the midterm vs the final exam?

Normal z-scores

Looks normal

```r
hist(nba$two_pp)
```



Still looks normal

```r
zscores <- (nba$two_pp - mean(nba$two_pp)) /
  sd(nba$two_pp)
hist(zscores)
```

Mean and SD

```
mean(zscores)

[1] 4.166e-16

sd(zscores)

[1] 1
```

The "blah e-16" is just R's way of saying zero.

Recall: Relationships

- Let $y_i = a + bx_i$ for $i = 1, 2, \dots, n$.
- $\bar{y} = a + b\bar{x}$.
- $\text{median}(y_1, \dots, y_n) = a + b\ \text{median}(x_1, \dots, x_n)$
- $\text{SD}(y) = |b|\text{SD}(x)$
- $\text{MAD}(y) = |b|\text{MAD}(x)$

You prove

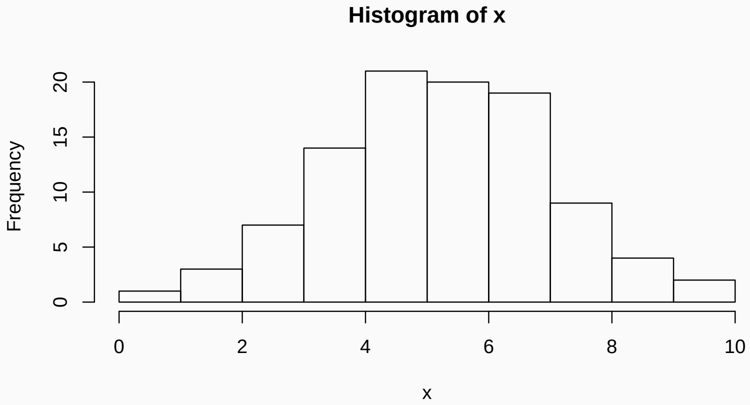
- **Claim**: Let $z_i = \frac{x_i - \bar{x}}{s_x}$ for $i = 1, \ldots, n$. Then $\bar{z} = 0$ and $s_z = 1$.

Property of Normal Distributions

- Actually, if we apply a linear transformation to a variable that has a normal distribution, then the resulting variable also has a normal distribution.
- Thus, if x is normal with mean μ and variance σ^2, then $z = \frac{x-\mu}{\sigma}$ is normal with mean 0 and variance 1.

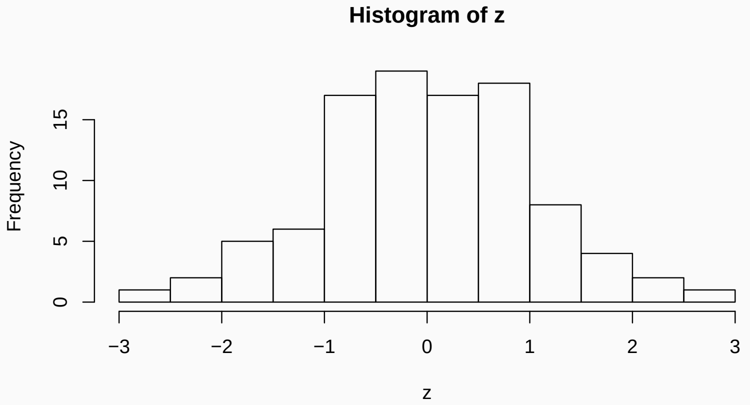
Normal z-scores

```r
x <- rnorm(n = 100, mean = 5, sd = 2)
hist(x)
```

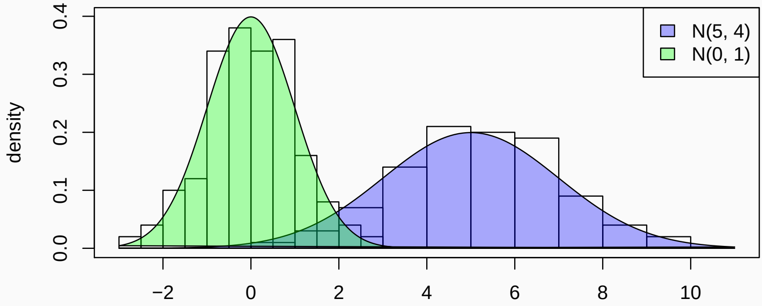


Normal z-scores

```
z <- (x - mean(x)) / sd(x)
hist(z)
```

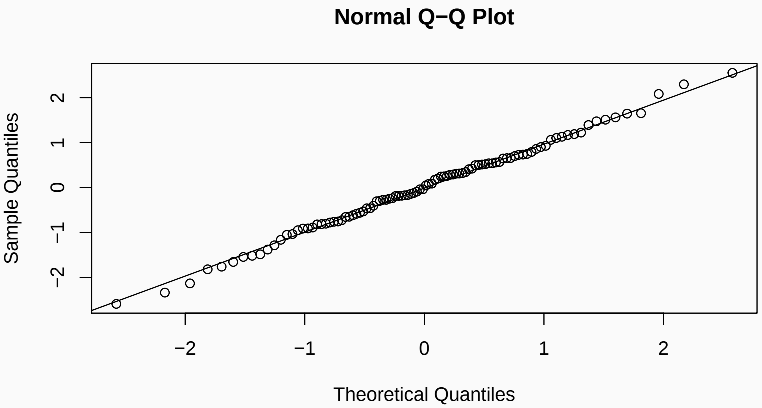


$N(5, 2^2)$ and $N(0, 1)$ on same plot



Check normality

```
qqnorm(z)
qqline(z)
```



Normal Probability Calculations

Approximations

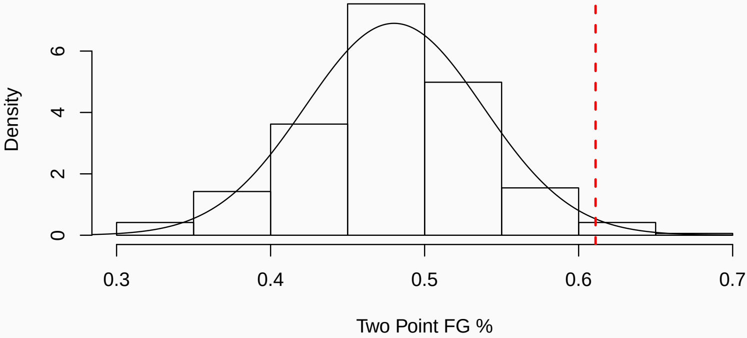
- We know LeBron's two-point field goal percentage. What percent of NBA players have a worse percentage?
- We could either calculate this out directly

```
lj2 <- nba$two_pp[nba$player == "LeBron James"]
sum(nba$two_pp < lj2) / length(nba$two_pp)
```

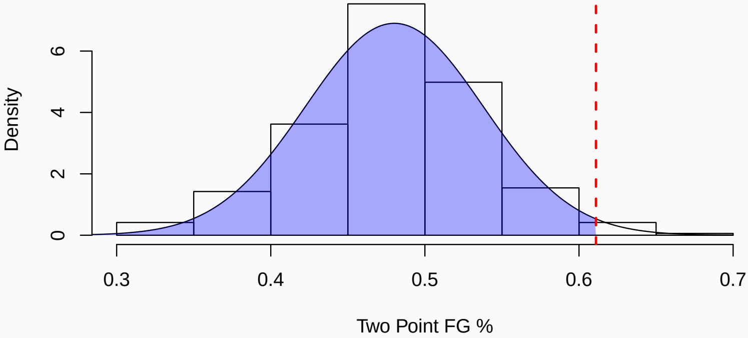
```
[1] 0.9881
```

- Or we could use a the normal distribution as an approximation.

Normal approximation



Area we want



Easy Way: use ‘pnorm‘

```
pnorm(q = lj2, mean = mean(nba$two_pp),
      sd = sd(nba$two_pp))
```

```
[1] 0.9882
```

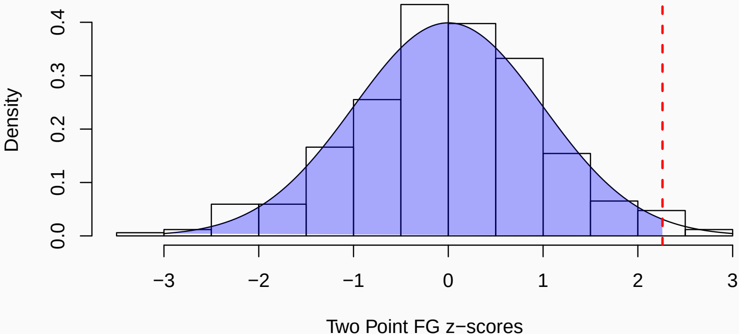
Pretty close to the observed frequency!

```
sum(nba$two_pp < lj2) / length(nba$two_pp)
```

```
[1] 0.9881
```

The Hard Way: Convert to z-scores and use a table

- Proportion of players who have a two-point FG% less than that of LeBron = proportion of players whose z-score is less than that of Lebron.
- Recall LeBron's z-score: $z_{lj} = 2.26$



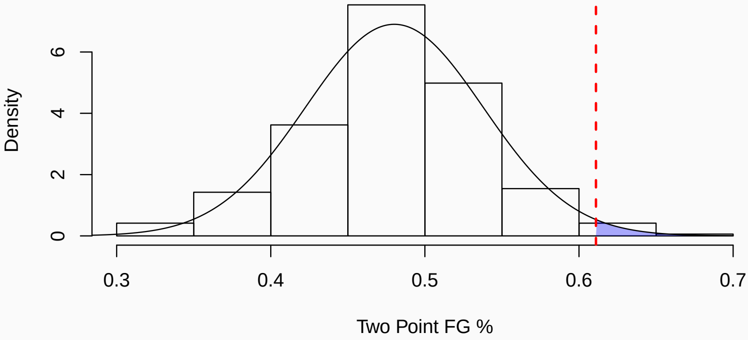
Table

- Want area to the left of 2.26 from a normal distribution with mean 0 and standard deviation 1.
- Look this up in Table B in DBC pp427-429.

	Second decimal place of Z								
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
					⋮				
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934
					⋮				

Another Problem

What about the Proportion of players who are better two-point field goal shooters than LeBron?



The Easy Way

```
1 - pnorm(q = lj2, mean = mean(nba$two_pp),
          sd = sd(nba$two_pp))
```

```
[1] 0.0118
```

```
pnorm(q = lj2, mean = mean(nba$two_pp),
      sd = sd(nba$two_pp),
      lower.tail = FALSE)
```

```
[1] 0.0118
```

```
sum(nba$two_pp >= lj2) / length(nba$two_pp)
```

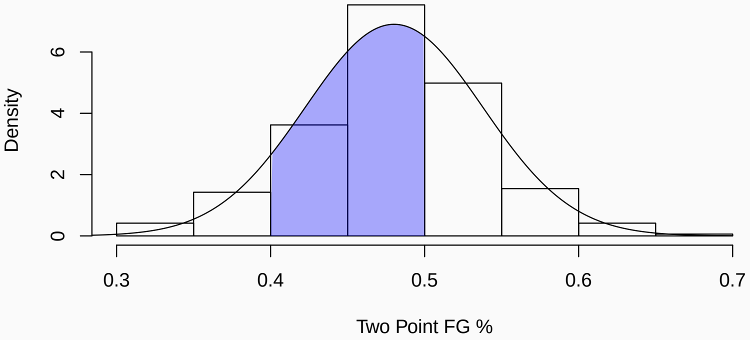
```
[1] 0.01187
```

The hard way

White Board

Another Problem

What proportion of NBA players shoot between 0.4 and 0.5 for two-point FG?



The Easy Way

```
less5 <- pnorm(0.5, mean = mean(nba$two_pp),
               sd = sd(nba$two_pp))
less4 <- pnorm(0.4, mean = mean(nba$two_pp),
               sd = sd(nba$two_pp))
less5 - less4
```

```
[1] 0.5515
```

```
sum(nba$two_pp >= 0.4 & nba$two_pp <= 0.5) /
  length(nba$two_pp)
```

```
[1] 0.5638
```

The hard way

White Board