Sampling Distributions

David Gerard

2017-09-18

Learning Objectives

- Statistics/parameters
- Sampling Distribution
- Sections 1.3.1, 1.3.2, 1.3.3, 4.1, 4.4 in DBC

Population and Sample

Recall: Population and Sample

population

A population is a set of cases (observational units) about which information is wanted.

sample

A sample is a subset of the population.

Examples

- We want to know demographic information of Americans so we randomly select a group of 50 Americans and ask them a bunch of questions. (sample? population?)
- We are interested in the quality of anchovies so we take 10 cans and taste them. (sample? population?)

Why sample?

- It is expensive/impossible to collect information on the whole population (when this is done it is called a census).
- Even when a census is performed, it is often less accurate than a well-designed sample (hard to collect information on everything, so this introduces biases into the observations you see).
- With a large enough sample, we can be pretty sure of the information we want on the population, making taking a census unnecessary.

Random Sampling

- Often, samples are collected randomly to remove bias.
- bias is where some cases are more likely to be in the sample than other cases.
- E.g. some political pollsters mostly call landlines, which biases the sample toward older individuals. What could be the issue here?

Statistics and Parameters

parameter

A parameter is a number that describes a population. It is usually unknown and what we want information on. People usually use greek letter μ, σ, ρ to represent parameters.

statistic

A statistic is a number that describes a sample. It is known and is used to estimate a population parameter. People usually use latin letters $\bar{x}, s, r$ to represent statistics.

Example

- We want to know the average height of U.S. males so we measure the average height of a sample of 50 U.S. males and came up with 5'11". (parameter? statistic?)

The sample mean

Recall: 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.

Recall: 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...
```

The inference problem

- Suppose I want to know the average total points of NBA players. However, I can only collect a sample of 5 players.

```
nsamp <- 5
samd <- sample(nba$pts, size = nsamp)
samd
```

```
[1]  709  479  130 1028  142
```

Point Estimate

Of course, we know the actual mean number of points μ because we have the entire population.

```
mean(nba$pts)

[1] 666.4
```

A good estimate might be the average of the sample $\bar{x}$

```
mean(samd)

[1] 497.6
```

Point Estimate

The sample average here is a point estimate of the population mean.

point estimate

A point estimate is a single number used to estimate a population parameter.

Aside

- How would you estimate the population median?
- How would you estimate the population standard deviation?

A different sample

However, since the sample was drawn at random, we could have obtained a different sample, and so a different point estimate.

```
samd <- sample(nba$pts, size = nsamp)
samd
```

```
[1]   94  419  435 1742 1025
```

```
mean(samd)
```

```
[1] 743
```

And another sample

```
samd <- sample(nba$pts, size = nsamp)
samd

[1] 1071  381  689  282  551

mean(samd)

[1] 594.8
```

And another sample

```
samd <- sample(nba$pts, size = nsamp)
samd
```

```
[1] 327  59 700 281 107
```

```
mean(samd)
```

```
[1] 294.8
```

And another sample

```
samd <- sample(nba$pts, size = nsamp)
samd

[1] 1002  425 1196  864  689

mean(samd)

[1] 835.2
```

Sampling distribution

- With every sample we are getting a different $\bar{x}$.
- We can ask what possible values $\bar{x}$ can take and how often it takes those values.
- That is, we can ask about $\bar{x}$'s *distribution*.

Sampling distribution

sampling distribution

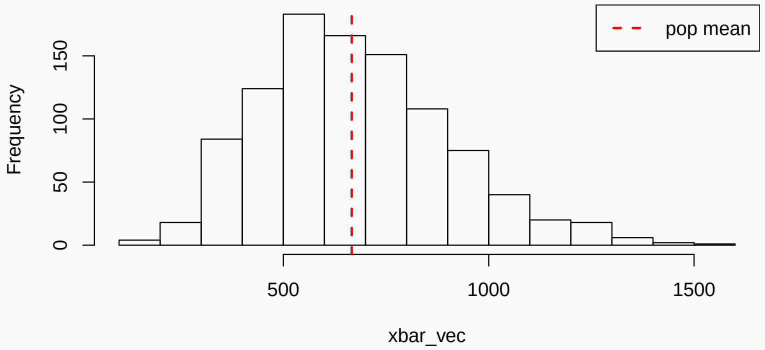
A sampling distribution is the distribution of a sample statistic.

Repeat sample 1000 times.

```
itermax <- 1000
xbar_vec <- rep(NA, itermax)
for (index in 1:itermax) {
  samd <- sample(nba$pts, size = nsamp)
  xbar_vec[index] <- mean(samd)
}
```

Plot the results

```
hist(xbar_vec, main = "")
abline(v = mean(nba$pts), lty = 2, col = 2, lwd = 2)
legend("topright", "pop mean", lty = 2, col = 2, lwd = 2)
```



The sampling distribution

- The sample mean has the correct center.
- There is a lot of variability about that center though.

```
sd(xbar_vec)
```

```
[1] 227.7
```

standard error

The standard deviation associated with a point estimate is called a standard error.

What if we have a bigger sample

```
nsamp <- 10
xbar10_vec <- rep(NA, itermax)
for (index in 1:itermax) {
  samd <- sample(nba$pts, size = nsamp)
  xbar10_vec[index] <- mean(samd)
}
sd(xbar10_vec)

[1] 156.3
```

What if we have a bigger sample

```
nsamp <- 50
xbar50_vec <- rep(NA, itermax)
for (index in 1:itermax) {
  samd <- sample(nba$pts, size = nsamp)
  xbar50_vec[index] <- mean(samd)
}
sd(xbar50_vec)
```

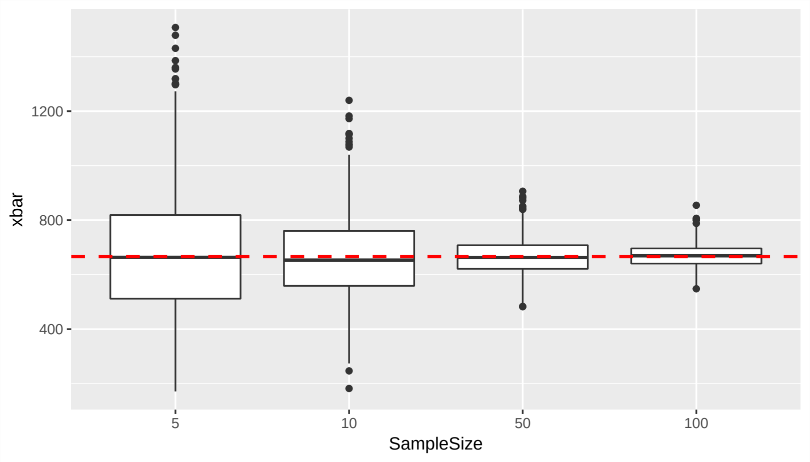
```
[1] 66.51
```

What if we have a bigger sample

```
nsamp <- 100
xbar100_vec <- rep(NA, itermax)
for (index in 1:itermax) {
  samd <- sample(nba$pts, size = nsamp)
  xbar100_vec[index] <- mean(samd)
}
sd(xbar100_vec)

[1] 42.61
```

Standard error decreases with larger sample sizes!



Dashed red line is population mean.

Standard error

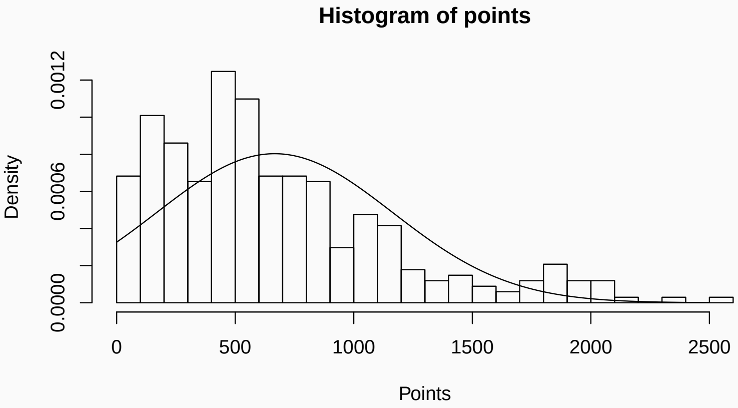
standard error

Given n independent observations from a population with standard deviation σ, the standard error of the sample mean is equal to

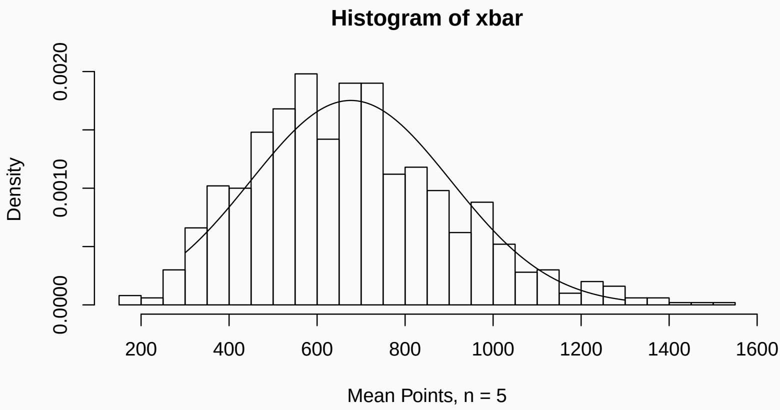
$$SE = \frac{\sigma}{\sqrt{n}}.$$

- Since σ is generally unknown, we estimate SE with $s/\sqrt{n}$, where s is the sample standard deviation.

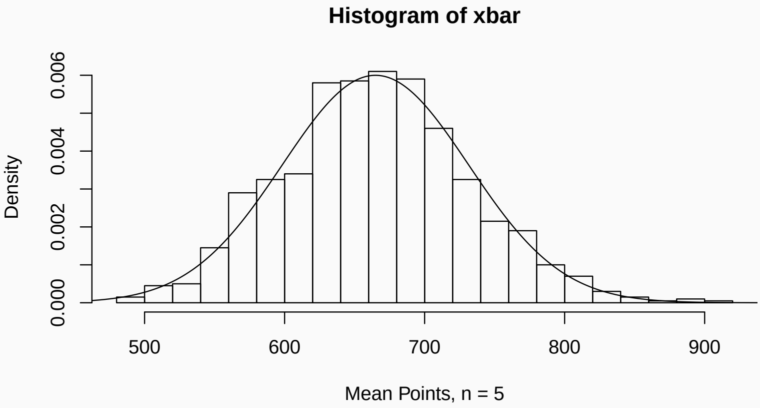
What happens as sample size increases?



What happens as sample size increases?



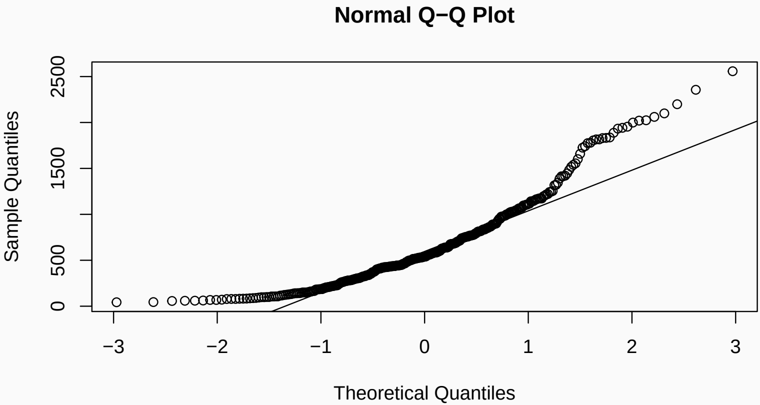
What happens as sample size increases?



Wat happens as the sample size increases?

$n = 1$

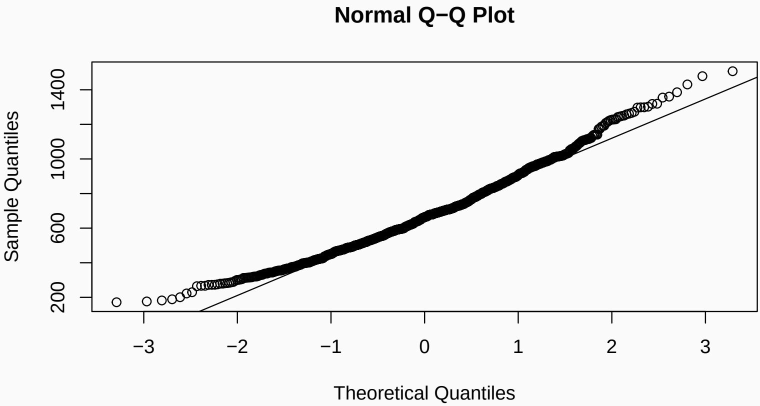
```
qqnorm(nba$pts)
qqline(nba$pts)
```



Wat happens as the sample size increases?

$n = 5$

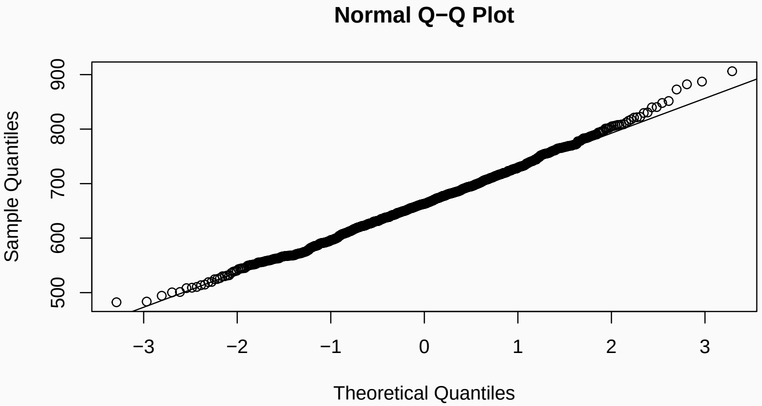
```
qqnorm(xbar_vec)
qqline(xbar_vec)
```



Wat happens as the sample size increases?

$n = 50$

```
qqnorm(xbar50_vec)
qqline(xbar50_vec)
```



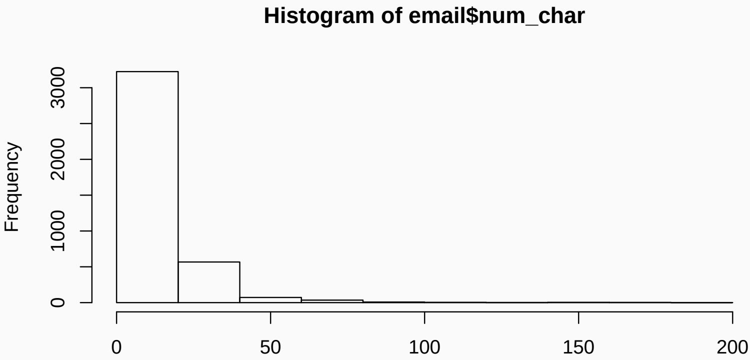
General result

- In general, sample means converge to a normal distribution as the sample size increases.
- Many other statistics do this as well (proportions, medians, standard devaitions).
- We will provide a heuristic proof of this result later.

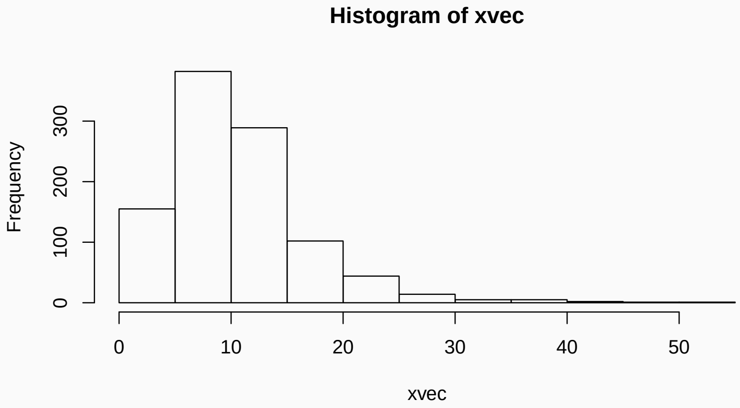
Skewed distributions

For highly skewed distributions, it takes more samples for normality to be a good approximation.

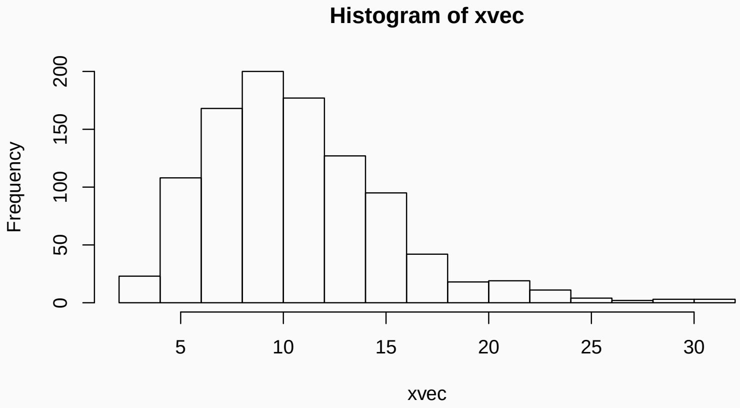
```
data(email, package = "openintro")
hist(email$num_char)
```



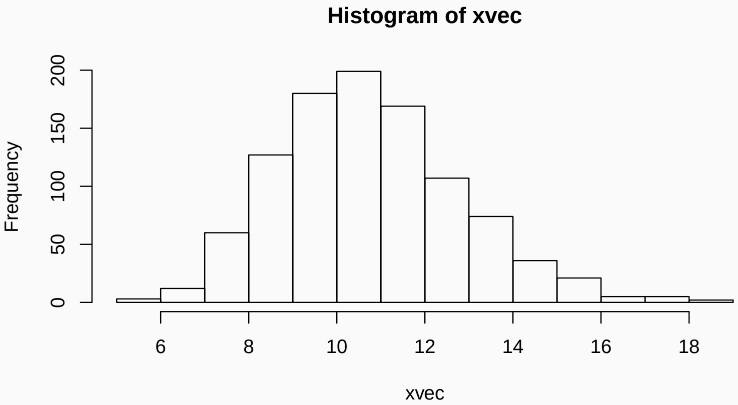
Skewed distributions, $n = 5$



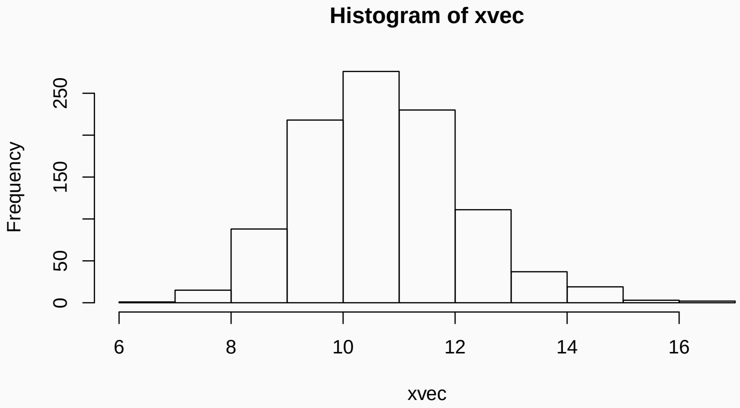
Skewed distributions, $n = 10$



Skewed distributions, $n = 50$



Skewed distributions, $n = 100$



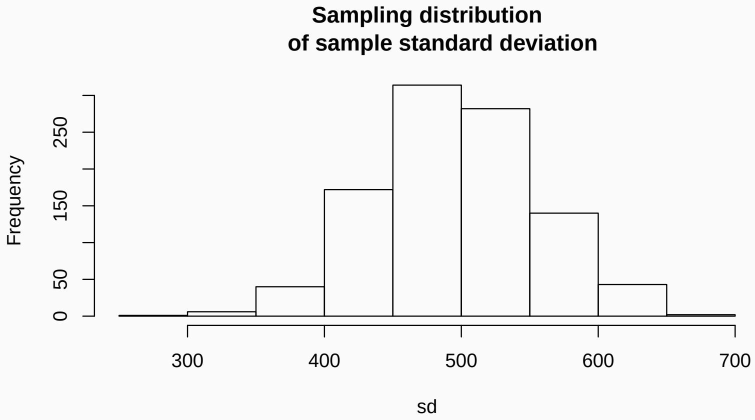
More sampling distributions

Every statistic has a sampling distribution

```
nsamp <- 50
sd_vec <- rep(NA, itermax)
for (index in 1:itermax) {
  samd <- sample(nba$pts, size = nsamp)
  sd_vec[index] <- sd(samd)
}
```

Every statistic has a sampling distribution

```
hist(sd_vec, main = "Sampling distribution
     of sample standard deviation",
     xlab = "sd")
```

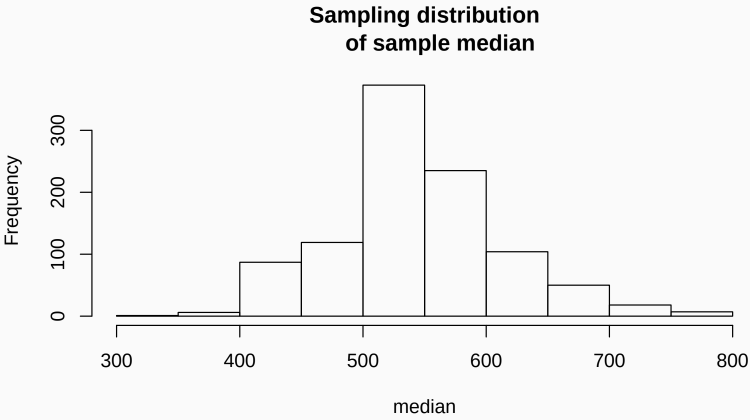


Every statistic has a sampling distribution

```
nsamp <- 50
med_vec <- rep(NA, itermax)
for (index in 1:itermax) {
  samd <- sample(nba$pts, size = nsamp)
  med_vec[index] <- median(samd)
}
```

Every statistic has a sampling distribution

```r
hist(med_vec, main = "Sampling distribution
     of sample median",
     xlab = "median")
```



Every statistic has a sampling distribution, but not all sampling distributions converge to a normal

```r
nsamp <- 50
max_vec <- rep(NA, itermax)
for (index in 1:itermax) {
  samd <- sample(nba$pts, size = nsamp)
  max_vec[index] <- max(samd)
}
```

Every statistic has a sampling distribution, but not all sampling distributions converge to a normal

```
hist(max_vec, main = "Sampling distribution
     of sample maximum",
     xlab = "max")
```