Probability and Statistics 101

Can we ever beat the Casino?

Objectives

- **Basic elements of Probability**

  The world is full of randomness. It is hard to predict what will exactly happen next. However, we can describe the randomness using probability. We will use a simple game to encapsulate the basic elements of probability: a sample space, events and probability.

- **Basic concepts of Statistics**

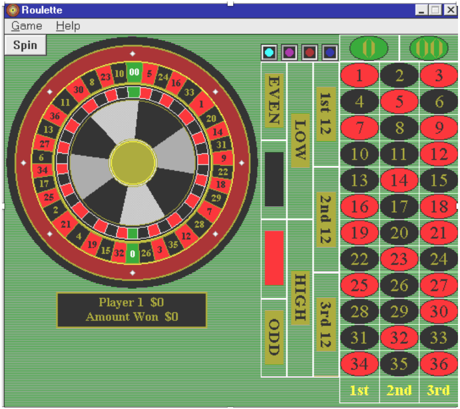
  We learn and infer the world using what we have observed.

- Gambling and probability

  Gambling shows that there has been an interest in quantifying the ideas of probability for millennia.

Table of Content

Roulette Game



- A wheel
  - 0, 00, 1, ..., 36
  - 18 numbers: red
  - 18 numbers: black
  - 0, 00: green
- A ball

Spin the wheel in one direction and spin the ball in the opposite direction. Observe where the ball lands.

Claim 1: A losing game

There are different ways to bet.

- Bet on one single number
- Bet on red or black

Claim 1

One will be for sure losing all the money in hands if playing the Roulette game MANY times.

Claim 2: An unfair game

I once went to a casino and played Red-Black games

- 100 times
- Each time bet \$1.00
- I lost \$28 at the end (Same as lost \$.28 on average)

Claim 2

The roulette table is not a fair one!

How to prove the claim?

We need the concept of **probability** and **statistics**.

Probablity

- In a roulette game, you can not predict where the ball is going to land. (Randomness)
- But ... We know the probability of events
  - Probability of seeing a 20 is $\frac{1}{38}$
  - Probability of seeing a red is $\frac{18}{38} = 0.47 < 1/2$
- What does Prob (seeing a 20) $= \frac{1}{38}$ mean?

Probability

What does Prob (seeing a 20) $= \frac{1}{38}$ mean?

One way: if one plays 1000 times, 20 will roughly appear $1000 \times \frac{1}{38} = 26$ times

Probability of a random event: a long term frequency.

Key elements:

- a sample space
- events
- probability

Random Variables (R.V.)

- A single number game (straight bet): Odds paid 35 to 1 (Put one dollar on a number (say 10) and you will win 35 (and get back your original \$1) if 10 appears; or you will lose \$1)
- Let X be the money won for one dollar bet, it is called a random variable.
- What are the possible values and corresponding prob?

$$X = 35 \quad \text{or} \quad X = -1$$

- Random variables are functions of the sample space.

Distributions

- The possible values together with their probabilities is called the distribution
  - If we win: $X = 35$ with prob $\frac{1}{38}$
  - If we lose: $X = -1$ with prob $1 - \frac{1}{38} = \frac{37}{38}$
- On average how much do you expect that we will win?

Expected Value

- On average how much do you expect that we will win?

$$E(X) = 35 \times \frac{1}{38} + (-1) \times \frac{37}{38}$$
$$= \frac{35}{38} - \frac{37}{38} = -\frac{2}{38} = -.0526$$

- Jargon: $-.0526$ is called the **expected** value of X. It is the weighted average of X and is denoted by $E(X)$.
- Question: What does -.0526 tell us?

Another game: Red-Black | Odds paid 1 to 1

- Put one dollar on one color, say red. If any of the red numbers appears you win \$1, otherwise you lose \$1
- Let Y be the money won for one dollar bet.
  - If we win: $Y = 1$ with prob $\frac{18}{38}$
  - If we lose: $Y = -1$ with prob $1 - \frac{18}{38} = \frac{20}{38}$
- The expected winning is now

$$E(X) = 1 \times \frac{18}{38} + (-1) \times \frac{20}{38} = -\frac{2}{38} = -.0526$$

- *This is same as the expected winning of one number game!!!!!*

Interpretation of Expected Value

- When we play Red-Black games on one dollar bet, we expect to win -0.0526, that is, on average we are going to lose 5.26 cents.
- Let us see what does -0.0526 mean.

  I was in Las Vegas not too long ago and I played Red-Black game 200 times. I only bet one dollar each time.

Interpretation of Expected Value

- Here is the summary of the 200 Red-Black games:

	Actual	Expected
Lost	105 times	$200 \times \frac{20}{38} = 105.3$
Won	95 times	$200 \times \frac{18}{38} = 94.7$
Average Winning	$\bar{Y}_{200} = \frac{Y_1 + \ldots + Y_{200}}{200} = (-105 + 95)/200 = -0.050$	-0.0526

Are you surprised to see this?

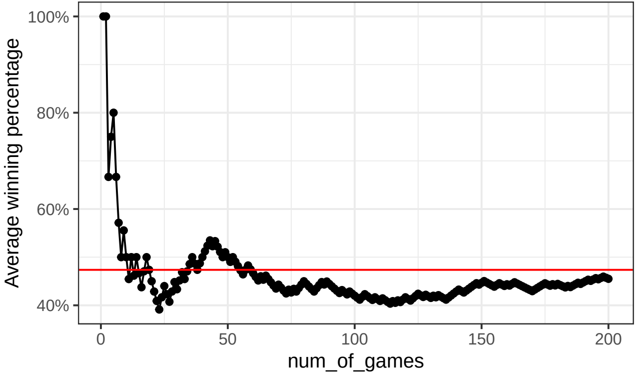
Behavior of $\bar{Y}_n$

```
n <- 200
win_prob <- 18/38
# winning event
win_vec <- rbinom(n,1,win_prob)
# cumulative mean = cumulative sum / cumulative number of games
cum_mean <- cumsum(win_vec)/(1:n)
# data.frame
cum_mean_df <- data.frame(num_of_games = 1:n,
                          cum_mean = cum_mean)
head(cum_mean_df)
```

```
##   num_of_games  cum_mean
## 1            1 0.0000000
## 2            2 0.5000000
## 3            3 0.3333333
## 4            4 0.2500000
## 5            5 0.4000000
## 6            6 0.5000000
```

Behavior of $\bar{Y}_n$

```
ggplot(cum_mean_df,
     aes(x = num_of_games, y = cum_mean)) +
geom_line() + geom_point() +
geom_hline(yintercept = win_prob, col = "red") +
# xlab("Number of games") +
ylab("Average winning percentage") +
scale_y_continuous(labels = percent) +
theme_bw()
```



Behavior of $\bar{Y}_n$

```r
# expected gain
expected_gain <- win_prob - (1-win_prob)

# if win: +1; if lose: -1
gain_vec <- win_vec*2-1

# sample() function:
# x: elements to choose; size: repeat how many times;
# replace: sample with replacement; prob: probability to choose each x
gain_vec <- sample(x = c(1,-1),
                size = n,
                replace = T,
                prob = c(win_prob, 1- win_prob))

gain_vec
```

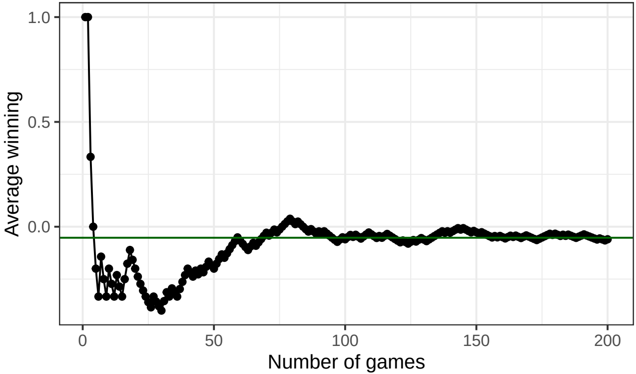
```
##    [1]  1  1 -1 -1 -1 -1  1 -1 -1  1 -1 -1  1 -1 -1  1  1  1 -1 -1 -1 -1 -1 -1 -1
##   [26] -1  1 -1 -1 -1  1  1 -1  1 -1 -1  1  1  1  1 -1 -1  1 -1  1 -1  1  1 -1 -1
##   [51]  1  1  1 -1  1  1  1  1  1 -1 -1 -1 -1  1  1 -1  1  1  1  1 -1  1  1 -1  1
##   [76]  1  1  1  1 -1 -1  1 -1 -1 -1 -1  1 -1 -1  1 -1  1 -1 -1 -1 -1 -1  1  1 -1
##  [101]  1  1 -1  1 -1 -1  1  1  1 -1 -1 -1  1 -1  1  1 -1 -1 -1 -1 -1  1 -1 -1  1
##  [126]  1 -1  1  1 -1 -1  1  1  1  1  1  1 -1  1 -1  1  1  1 -1  1 -1 -1 -1  1 -1
##  [151] -1  1 -1 -1 -1 -1  1 -1  1 -1 -1  1  1 -1  1 -1 -1  1  1 -1 -1 -1 -1  1  1
##  [176]  1  1  1 -1  1 -1 -1  1 -1  1 -1 -1 -1  1  1  1 -1 -1 -1 -1 -1  1 -1 -1  1
```

Behavior of $\bar{Y}_n$

```r
# average gain
ave_gain <- cumsum(gain_vec)/1:n

# data.frame
cum_gain_df <- data.frame(num_of_games = 1:n,
                          ave_gain = ave_gain)
```

```r
# plot
ggplot(cum_gain_df,
     aes(x = num_of_games, y = ave_gain)) +
geom_line() + geom_point() +
geom_hline(yintercept = expected_gain, col = "darkgreen") +
xlab("Number of games") +
ylab("Average winning") +
theme_bw()
```

Behavior of $\bar{Y}_n$



Law of Large Numbers

- The expected winning for Red-Black game is -0.0526
- Long term Average $\approx$ expected value

$$\bar{Y}_n \to \mu \text{ or } E(Y) \text{ (Expected value)}$$

Which game is better?

- The expected winning for Red-Black game is -0.0526
- Recall that the expected winning for Single number bet is also -0.0526
- Both games have the same expected values.

Which game should we play to make money?

Risk measurement: Variance

HOW? Little long stories!

Variability: Variance

- X=winning on a single number bet: It can be 35 or -1 with prob 1/38 or 37/38. The expected winning is -0.0526
- Variance: the expected squared difference of the winning from the expected winning $=E(X-\mu)^2 = \sigma^2 = VAR(X)$:

$$\sigma_X^2 = (35-(-0.0526))^2 \times \frac{1}{38} + (-1-(-0.0526))^2 \times \frac{37}{38} = 33.208$$

- Standard Deviation:

$$\sqrt{\sigma_X^2} = \sqrt{33.208} = 5.76$$

Notice: Expected values and Variances are theoretical quantities. They are different from sample means and sample variances.

Standard Deviation for Y, the winning for Red-Black game?

- Y takes value 1 and -1 with prob. 18/38 and 20/38
- $$Var(Y) = (1-(-0.0526))^2 \times \frac{18}{38} + (-1-(-0.0526))^2 \times \frac{20}{38} = 0.997$$
- $$\sigma_Y = \sqrt{0.997} = 0.998$$
- The variability of winning from a single number game (SD=5.76) is much larger than that of Red-Black (SD=0.998)
- How do Variances help us to determine which game to play?

Behavior of the average winning

(Sample of size 10, 100, 10000 vs. the population)

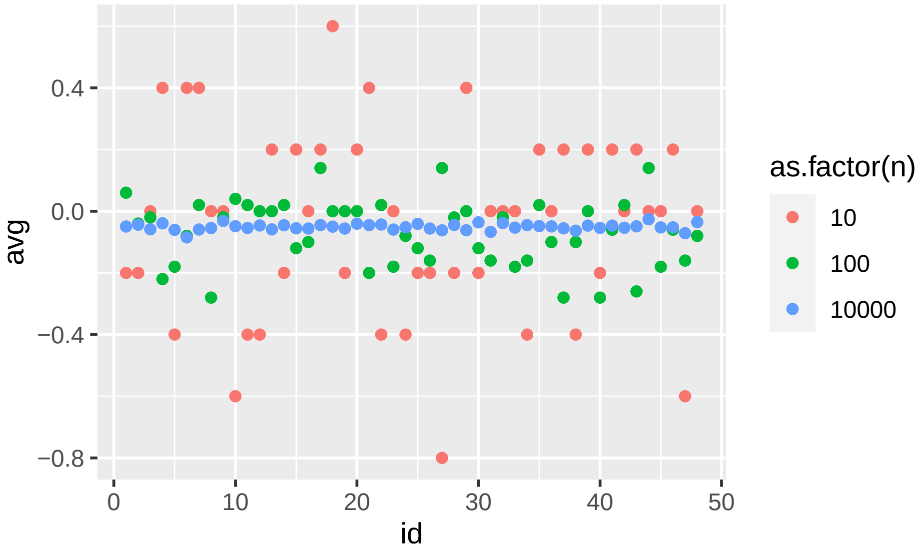
We all play Red-Black game, bet one dollar each time

- $\bar{Y}_{10}$, each person play 10 times 48 of us
- $\bar{Y}_{100}$, each person play 100 times 48 of us
- $\bar{Y}_{10,000}$, each person play 10,000 times 48 of us

Behavior of the average winning

```r
n_times <- 48
win_prob <- 18/38

# create a data frame
## 10 times
set.seed(1)
avg_winning_df_10 <-
  data.frame(id = 1:n_times,
             n = 10,
             num_win = rbinom(n_times, 10, win_prob))
## 100 times
avg_winning_df_100 <-
  data.frame(id = 1:n_times,
             n = 100,
             num_win = rbinom(n_times, 100, win_prob))

# 10000 times
avg_winning_df_10000 <-
  data.frame(id = 1:n_times,
             n = 10000,
             num_win = rbinom(n_times, 10000, win_prob))

avg_winning_df <- rbind(avg_winning_df_10, avg_winning_df_100, avg_winning_df_10000)

avg_winning_df <-
  avg_winning_df %>%
  mutate(avg = (num_win - (n-num_win))/n )

## another way
# times <- c(10, 100, 10000)
# ns <- rep(times, each = n_times)
# avg_winning_df <-
#   data.frame(id = rep(1:n_times, 3),
#              n = ns,
#              num_win = unlist(lapply(times,
```

Behavior of the average winning



Behavior of the average winning

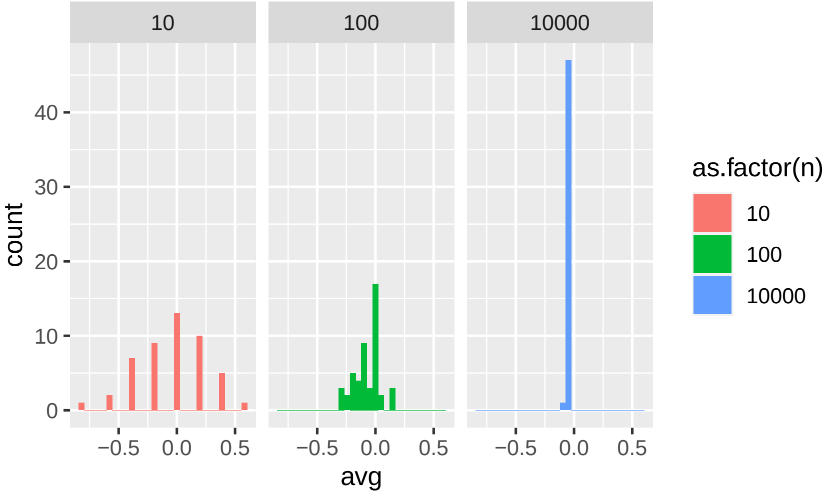
```
avg_winning_df %>%
  group_by(n) %>%
  summarize(mean = mean(avg),
            sd = sd(avg),
            total = sum(avg*n))
```

```
## # A tibble: 3 x 4
##       n    mean     sd  total
##   <dbl>   <dbl>  <dbl>  <dbl>
## 1    10 -0.0417 0.303     -20
## 2   100 -0.0704 0.109    -338
## 3 10000 -0.0510 0.0104 -24466
```

Behavior of the average winning

```
ggplot(avg_winning_df, aes(x = avg, fill = as.factor(n))) +
  geom_histogram() +
  facet_wrap(~n, nrow = 1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Central Limit Theorem (CLT)

- When a large number of games are played
  - The average amount each person wins (lost in this case) tends to be close to the center = "expectation'' (-0.0526)
  - The distribution is also approximately a bell curve!
- The Central Limit Theorem
  - $\bar{Y}_n$ has a normal distribution
  - $E[\bar{Y}_n] = \mu/n$
  - $Var(\bar{Y}_n) = \sigma^2/n$
- Almost for sure each one of us will lose all the money if we keep playing!

Single number games

What about instead we have all played single number games?

Single number game

```r
# winning probability
win_prob = 1/38
# number of game
n_times <- 48
# number of trials each game
times <- c(10, 100, 10000)
ns <- rep(times, each = n_times)
# number of win
num_win <- c(sapply(times,
                  function(trial) rbinom(n_times, trial, win_prob)))

avg_winning_df <- data.frame(id = rep(1:n_times, 3),
                             n = ns,
                             num_win = num_win)

avg_winning_df <-
  avg_winning_df %>%
  mutate(avg = (num_win*35 - (n-num_win))/n )
```

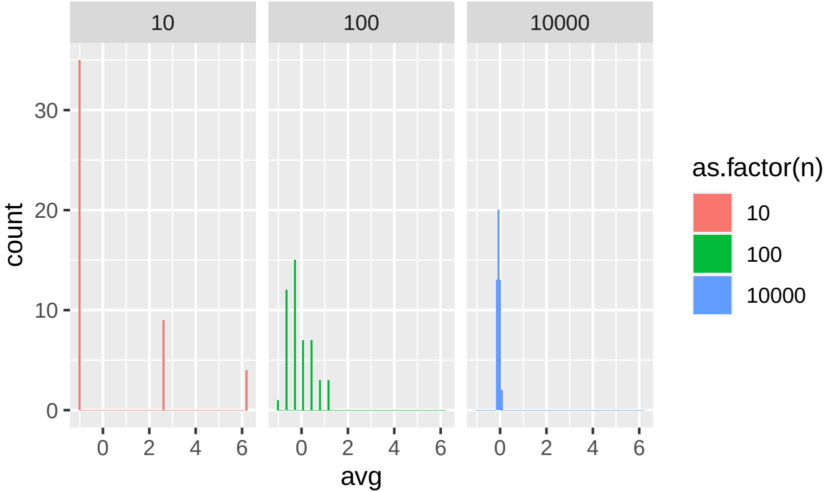
Single number game

```
avg_winning_df %>%
  group_by(n) %>%
  summarize(mean = mean(avg),
            sd = sd(avg),
            total = sum(avg*n))
```

```
## # A tibble: 3 x 4
##         n    mean     sd  total
##     <dbl>   <dbl>  <dbl>  <dbl>
## 1     10  0.275  2.29      132
## 2    100 -0.07   0.550    -336
## 3 10000 -0.0674 0.0568 -32340
```

Single number game

```
ggplot(avg_winning_df, aes(x = avg, fill = as.factor(n))) +
  geom_histogram(bins = 100) +
  facet_wrap(~n, nrow = 1)
```



Summary of two games: Single number vs Red-Black

- The expected winning is same: -.0526 on one dollar
- Single number:
  - One may have chance to win large amount
  - BUT one may also lose a lot
  - On average you come out the same as Red-Black
- Red-Black:
  - Much more conservative
  - If you want to kill time you may choose this game

After all: Almost for sure to lose money if one plays many times

Take away:

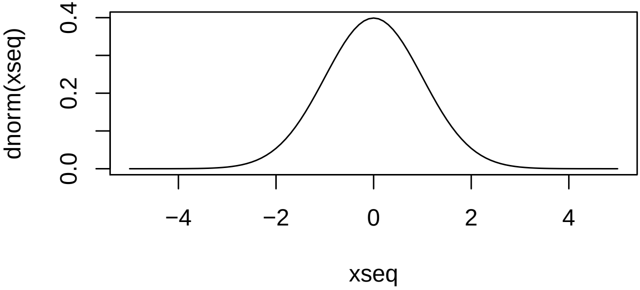
- You can not tell for sure what will happen for a random event.
- Probability tells us on average how often the event will occur.
- A random number changes
  - The center: expected value
  - The spread: standard deviation
- An average of random sample follows a bell curve
  - It tends to the expected value
  - The variability is much smaller when sample size is larger

Normal Random Variable

X = value drawn randomly from a normal population with mean μ and standard deviation σ.

- Often abbreviated as $X \sim N(\mu, \sigma^2)$.
- Density:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x-\mu)^2}{2\sigma^2}$$



The Standard Normal Variable Z

- $\mu = 0$ and $\sigma = 1$
- Example: find

$$P(-1 \leq Z \leq 1) = P(Z \leq 1) - P(Z < -1) = .842 - .159 \approx 68\%$$

$$P(-1.96 \leq Z \leq 1.96) = .95$$

$$P(-3 \leq Z \leq 3) \approx 1$$

Are those numbers familiar?

A Normal Variable X

- If $X \sim N(\mu, \sigma^2)$, let $Z = \frac{x-\mu}{\sigma}$, then $Z \sim N(0,1)$
- So
$$P(a \leq X \leq b) = P(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma})$$
-
$$P(\mu - 1\sigma \leq X \leq \mu + 1\sigma) = P(-1 \leq Z \leq 1) = 68\%$$
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = P(-2 \leq Z \leq 2) = 95\%$$
$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = P(-3 \leq Z \leq 3) = 100\%$$

Distribution, mean and variance of $\bar{Y}_n$

Example: If we play Red and Black games 100 times, we agree that the average winning $\bar{Y}_{100}$ follows a normal distribution with mean being

$$E(\bar{Y}_{100}) = \mu = -.0526$$

and a variance of

$$Var(\bar{Y}_{100}) = 0.997/100 \approx 0.01$$

$$\sigma_{\bar{Y}_{100}} = \sqrt{0.01} = .1$$

So

$$\bar{Y}_{100} \sim N(-.0526, 0.01)$$

Distribution, mean and variance of $\bar{X}_n$

Example: If we play a single number game 100 times, we agree that the average winning $\bar{X}_{100}$ follows a normal distribution with mean being

$$E(\bar{X}_{100}) = \mu = -.0526$$

and a variance of

$$\sigma_{\bar{X}_{100}} = 5.76/\sqrt{100} = .576$$

So

$$\bar{X}_{100} \sim N(-.0526, 0.576^2)$$

Comparison of two games

- 95% of time
  - $\bar{Y}_{100}$ will be within $-.0526 \pm 2 \times .1 = (-.25, .147)$
  - $\bar{X}_{100}$ will be within $-.0526 \pm 2 \times .576 = (-1.2, 1.09)$
- The chance for $\bar{Y}_{100} > .147$ is same as $\bar{X}_{100} > 1.09$, being 2.5%

Again, which game will you play?

More detailed calculations:

We can also find out:

a) Prob (positive winning)=Prob($\bar{Y}_{100} > 0$)
b) Prob (losing money)=Prob($\bar{Y}_{100} \leq 0$)
c) Prob($-.2 \leq \bar{Y}_{100} \leq -.1$)

Red and Black games 100 times

Recall $\bar{Y}_{100} \sim N(-.0526, 0.01)$.

a) Prob (positive winning)=Prob($\bar{Y}_{100} > 0$)

$$P(\bar{Y}_{100} \geq 0) = P\left(Z \geq \frac{0-(-.0526)}{.1}\right)$$
$$= P\left(Z \geq .526\right) = .3$$

```
pnorm(.526, lower.tail = F)
```

```
## [1] 0.2994441
```

```
# pnorm(0, mean = -.0526, sd = .1, lower.tail = F)
```

Red and Black games 100 times

(b) Prob (losing money)=Prob($\bar{Y}_{100} \leq 0$) = 1- Prob($\bar{Y}_{100} > 0$) =1-.3=.7

On average the chance to lose money is 70%.

c) Prob($-.2 \leq \bar{Y}_{100} \leq -.1$)

$$P(-.2 \leq \bar{Y}_{100} \leq -.1) = P\left(\frac{-.2-(-.0526)}{.1} \leq Z \leq \frac{-.1-(-.0526)}{.1}\right)$$
$$= P\left(-1.474 \leq Z \leq -.474\right) = .32 - .07 = .25$$

```
pnorm(-.474) - pnorm(-1.474)
```

```
## [1] 0.2475092
```

The chance of loosing between 10 and 20 cents on average is 25%

Single number game 100 times

Recall $\bar{X}_{100} \sim N(-.0526, 0.576^2)$.

Prob(losing money):

$$\begin{aligned} P(\bar{X}_{100} < 0) &= P\left(Z < \frac{0 - (-.0526)}{.576}\right) \\ &= P\left(Z \geq .0913\right) = .536 \end{aligned}$$

On average the chance to lose money is 53.6%!

Is the Casino being honest?

Goal: Estimate the mean winning or test if the mean is -.056

How?

Data: Gather data by playing Red-Black game 100 times, we have observed:

```
set.seed(1)
n <- 100
win_vec <- rbinom(n, 1, 18/38)
gain_vec <- win_vec*2-1
data.frame(mean = mean(gain_vec),
           sd = sd(gain_vec),
           upper = mean(gain_vec) + 1.96*sd(gain_vec)/sqrt(n),
           lower = mean(gain_vec) - 1.96*sd(gain_vec)/sqrt(n)
)
```

```
##    mean       sd     upper      lower
## 1 -0.06 1.003227 0.1366325 -0.2566325
```

Question: Is -0.06 by chance?

95% Confidence interval

$\bar{X}$ has a normal distribution with μ and $sd = \frac{\sigma}{\sqrt{100}} = \frac{.998}{10} \approx .1$

Which means 95% of time

$$|\bar{X} - \mu| < 1.96 \times .1$$

This is same to say 95% time the mean μ should be in

$$\bar{X} \pm 1.96 \times \frac{\sigma}{\sqrt{100}} = (\bar{X} - .2, \bar{X} + .2)$$

Apply to our data, we have a 95% confidence interval (z):

$$-.28 \pm 2 \times .1 = (-.48, -.08)$$

Conclusion: The roulette is not fair. 95% CI does not contain -.0526.

t-Confidence interval

- σ is not known either, we estimate σ by $s = .965$
- We will have a t-interval:

$$\bar{X} \pm t_{df} \times \frac{s}{\sqrt{100}} = -.28 \pm 1.98 \times \frac{.965}{\sqrt{100}} = (-.471, -.089)$$

- We have the same conclusion that the wheel is not a fair one since the true mean -.0526 in not in the interval.
- t intervals are wider than z intervals

Hypotheses testing

- We may ask is it possible that$\mu = -.0526$?
- $H_0 : \mu = -.0526 \quad \text{vs.} \quad H_1 : \mu \neq -.0526$
- Testing statistics

$$t = \frac{\bar{X} - (-.0526)}{s/\sqrt{n}} = \frac{-.28 + .0526}{.965/sqrt100} = -.2357$$

- p-value $= P(|T| > 2.357) = .0204$ if $\mu = -.0526$
- Conclusion: Since p-value is so small, we reject H_0.

Bernoulli Distribution

The success of each bet X of the single number game or the Red-Black game follows a Bernoulli distribution. Denote success as 1.

- Single number game

$$X = \begin{cases} 1 & \text{with probability (w.p.) } 1/38 \\ 0 & \text{w.p. } 37/38 \end{cases}$$

- Red-Black game

$$X = \begin{cases} 1 & \text{w.p. } 18/38, \\ 0 & \text{w.p. } 20/38 \end{cases}$$

Bernoulli Distribution

Simulate 100 trials. Use `rbinom()` to generate random samples.

```
##   [1] 1 0 0 1 1 0 0 0 1 1 1 1 0 0 0 0 1 0 0 1 1 0 0 0 1 0 0 0 0
##  [38] 0 1 0 1 1 0 0 1 0 0 1 0 1 1 1 0 0 0 0 1 0 0 0 0 1 0 1 1 0
##  [75] 0 1 1 1 1 1 0 0 1 0 1 0 1 1 1 1 1 0 0 1 0 1 0 1 0 1
```

Binomial Distribution

If we bet 100 times, or say we draw 100 samples from the Bernoulli distribution, the total number of success among these 100 times Y follow binomial distribution.

$$Y \sim Binomial(n, p)$$

where n is the total number of trials and p is the probability of success of each trial.

- Single number game

$$X = Binomial(100, 1/38)$$

- Red-Black game

$$X = Binomial(100, 18/38)$$

Binomial Distribution

The probability of success k times among 100 trials is

$$Prob(Y = k) = \left(\begin{array}{c} 100 \\ k \end{array} \right) p^k (1 - p)^{n-k}$$

```
## [1] 0.7349765
```

Binomial Distribution

Simulate total number of success among 100 trials. Use `rbinom()` to generate random samples.

```
## [1] 50
```

Covariances: $Cov(X_R, X_B)$

- $X_R =$ Winning over one dollar bet on Red
- $X_B =$ Winning over one dollar bet on Black
- X_R and X_B are related: if $X_R = 1$, then $X_B = -1$
- We use covariance to measure the relationship

$$COV(X_R, X_B) = E(X_R - E(X_R)(X_B - E(X_B))$$

$$COV(X_R, X_B) = -.8975$$

- Or Correlation

$$\rho = \frac{COV(X_R, X_B)}{SD(X_R)SD(X_B)} = \frac{-.8975}{.998 \times 998} = -.9011$$

Correlation

$$\rho = \frac{COV(X_R, X_B)}{SD(X_R)SD(X_B)} = \frac{-.8975}{.998 \times 998} = -.9011$$

- Correlation captures linear relationship between XR and XB
- $-1 < \rho < 1$
- The larger $|\rho|$ is, the stronger of the relationship
- The sign of ρ reflects the direction of associations

$E(X_R + X_B)$ and $VAR(X_R + X_B)$

- $E(X_R + X_B) = E(X_R) + E(X_B)$
- $VAR(X_R + X_B) = VAR(X_R) + VAR(X_B) + 2COV(X_R, X_B)$
- $VAR(aX_R + bX_B) = a^2 VAR(X_R) + b^2 VAR(X_B) + 2abCOV(X_R, X_B)$
- If X and Y are independent $COV(X, Y) = 0$
  $VAR(aX + bY) = a^2 VAR(X) + b^2 VAR(Y)$
- That is why $Var(\bar{X}_n) = \frac{\sigma^2}{n}$