SURVIVOR

How to Become a Winner

What is Survivor?

Outwit, Outplay, Outlast to become Sole Survivor

In a complex social game, 18 players compete on a remote island in Fuji in physically and mentally challenging games as they vote people off the island one by one for the chance to win $1 million dollars and the title of Sole Survivor

The Variables

There has been much speculation on what makes a contestant a threat to win the game. Our group has chosen to look at four different variables that are the most commonly talked about.

- Rate of success in challenges
- If you are a returning player
- Age
- Personality type

Rate of Success In Challenges

Hypothesis

Does the rate of change affect your chances of winning?
Null Hypothesis: the rate of challenge wins has no effect on your chances of winning
Alternative Hypothesis: the rate of challenge wins has an effect on your chances of winning

Creating the Data Frame

We first created a table that had number of challenge wins, number of challenges competed in, and the percentage of challenges won:

````r
```{r}
# number of challenge wins per castaway
challenge_stats <- challenge_results |>
  group_by(castaway_id) |>
  summarise(
    won = sum(result == "Won"),
    total_challenges = n(),
    percentage_won = (won/total_challenges)
  )
glimpse(challenge_stats)
```
````

Next, we created a table to of the data of the contestants who did win their season:

````r
```{r}
df <- season_summary %>%
     left_join(challenge_results, by = c("winner_id" = "castaway_id"))
```

#df table has episode summaries for season winners

```{r}
# number of challenge wins per winning castaway
winners_challenge_stats <- df |>
  group_by(version_season.x) |>
  summarise(
    won = sum(result == "Won"),
    total_challenges = n(),
    percentage_won = (won/total_challenges)
  )
glimpse(winners_challenge_stats)
```
````

````
Rows: 60
Columns: 4
$ version_season.x <chr> "AU01", "AU02", "AU03", "AU04", "AU05", "AU06", "AU07", "NZ01", "…
$ won              <int> 7, 14, 18, 11, 30, NA, 19, 6, 12, 8, 8, 10, 9, 10, 16, 13, 14, 11…
$ total_challenges <int> 33, 42, 39, 39, 63, 38, 51, 20, 24, 24, 23, 24, 23, 26, 29, 26, 3…
$ percentage_won   <dbl> 0.2121212, 0.3333333, 0.4615385, 0.2820513, 0.4761905, NA, 0.3725…
````

We then filtered this table to isolate contestants who did not win their season:

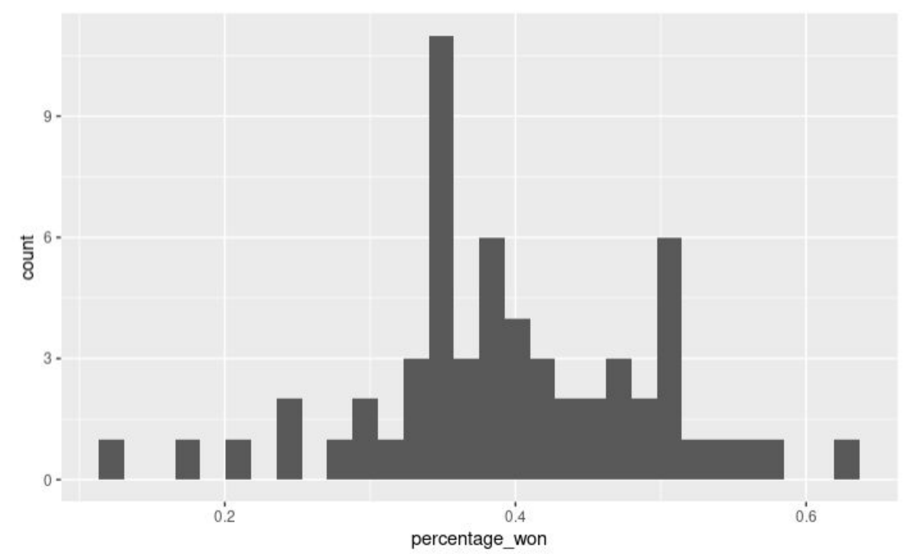
````r
```{r}
winner_list <- df %>%
  select(winner_id)

glimpse(winner_list)
```

```{r}
not_winners_challenge_stats <- challenge_stats %>%
  filter(!castaway_id %in% c(winner_list$winner_id))
glimpse(not_winners_challenge_stats)
```
````

````
Rows: 899
Columns: 4
$ castaway_id      <chr> "AU0001", "AU0002", "AU0003", "AU0004", "AU0005", "AU0006", "AU0…
$ won              <int> 0, 1, 1, 4, 5, 7, 5, 5, 7, 7, 15, 7, 7, 28, 17, 7, 16, 34, NA, 1…
$ total_challenges <int> 1, 2, 4, 5, 8, 10, 13, 14, 16, 17, 38, 19, 20, 45, 23, 25, 26, 6…
$ percentage_won   <dbl> 0.0000000, 0.5000000, 0.2500000, 0.8000000, 0.6250000, 0.7000000…
````

Histograms

Winners

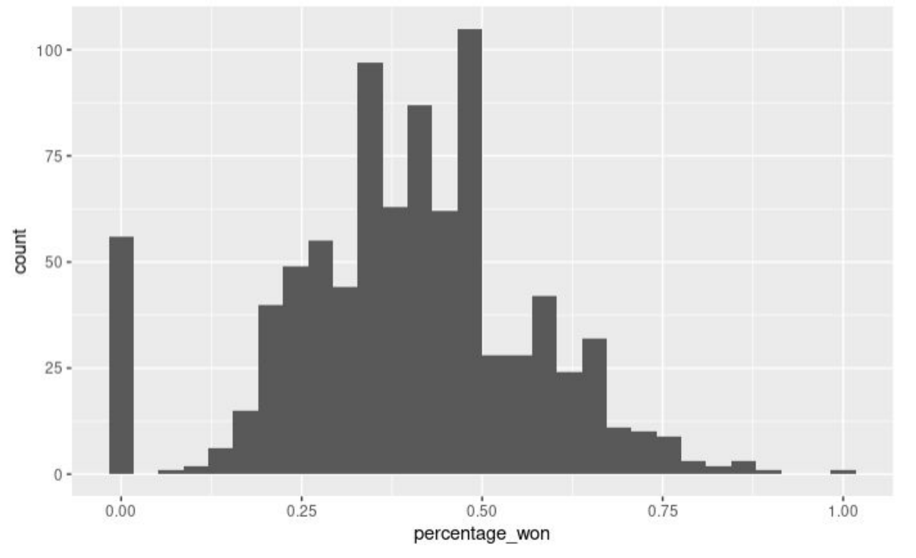
```{r}
# percentage of challenges won for winners histogram
ggplot(data = winners_challenge_stats, aes(x = percentage_won)) + geom_histogram()
```



Mean rate of success: 39.5174%

Non Winners

```{r}
# percentage of challenges won histogram (non winners)
ggplot(data = not_winners_challenge_stats, aes(x = percentage_won)) + geom_histogram()
```

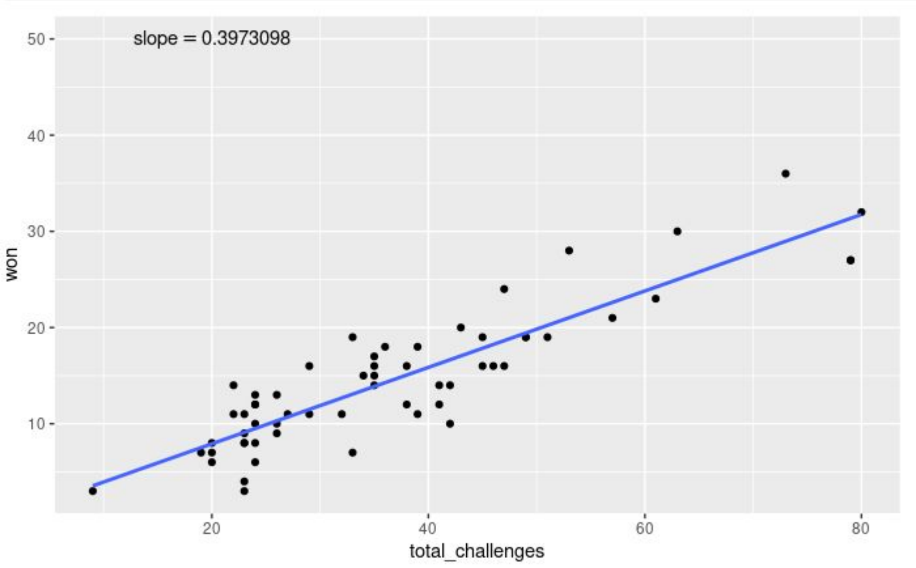


Mean rate of success: 39.4871%

Scatterplots

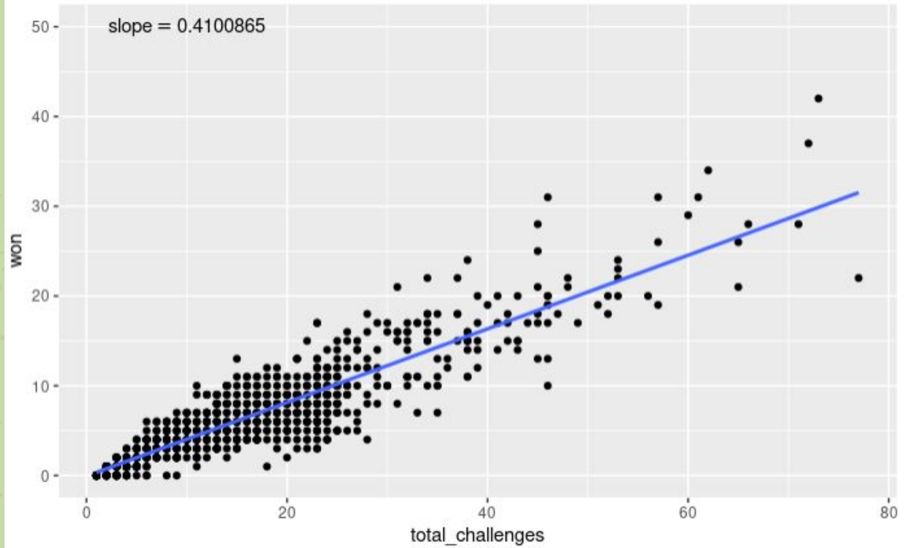
Winners

````
```{r}
# number of challenge wins versus number of challenges competed in (per castaway)
ggplot(data = winners_challenge_stats, aes(x = total_challenges, y = won)) + geom_point() +
geom_smooth(method=lm, se=FALSE) + annotate("text",x=20, y=50, label = (paste0("slope==",
coef(lm(winners_challenge_stats$won~winners_challenge_stats$total_challenges))[2])), parse=TRUE)
```
````



Non Winners

````
```{r}
# number of challenge wins versus number of challenges competed in (per non winning castaway)
ggplot(data = not_winners_challenge_stats, aes(x = total_challenges, y = won)) + geom_point() +
geom_smooth(method=lm, se=FALSE) + annotate("text",x=10, y=50, label = (paste0("slope==",
coef(lm(not_winners_challenge_stats$won~not_winners_challenge_stats$total_challenges))[2])),
parse=TRUE)
```
````



Hypothesis Testing

Does the rate of change of challenge wins affect your chances of winning?

I used an inference for two proportions test and created a confidence interval to check for error

```
```{r}
pw = mean(winners_challenge_stats$percentage_won, trim = 0, na.rm = TRUE)
nw = 60
```

```{r}
pt = mean(not_winners_challenge_stats$percentage_won, trim = 0, na.rm = TRUE)
nt = 899
```

```{r}
p_pool = (nw*pw + nt*pt)/(nw+nt)
```

```{r}
SE = sqrt(p_pool*(1 - p_pool)*((1/nw) + (1/nt)))
```

```{r}
z = ((pw-pt)-0)/SE
```

```{r}
p_value = 2*(1-pnorm(z,0,1))
```
```

```
```{r}
SE_CI = sqrt(((pw*(1-pw))/nw) + ((pt*(1-pt))/nt))
```

```{r}
z_star = qnorm(.95 + (.05/2), 0, 1)
```

```{r}
ME = z_star*SE_CI
```

```{r}
CI_lower = pw-pt - ME
```

```{r}
CI_upper = pw-pt + ME
```
```

Validation (Bootstrapping)

With an alpha of 0.05 used, the p_value found is greater than the alpha (0.569 > 0.05). Thus, we do not have sufficient evidence to reject the null hypothesis which states there is no correlation between number of challenges won and the probability of winning the season.

```
# library for bootstrap
library(boot)

#function to calculate correlation between percentage won and winning the show
bootstrap_func <- function(data, indices) {
  subset <- data[indices, ]
  cor(subset$percentage_won, subset$winner)
}

#no correlation between rate of challenge wins and the probablility of winning the season
null_hypothesis = 0

set.seed(321)
#use boot() from library to run bootstrap test for 1000 runs/samples
boot_result <- boot(boostrap_df,bootstrap_func, R=1000)

p_value <- mean(boot_result$t >= null_hypothesis)
```

```
> p_value
[1] 0.569
```

Conclusion

I found a p-value of .996 which meant there was not enough evidence to reject the null hypothesis.
My Confidence interval also supported this.
Therefore, the rate of challenge wins does not impact your chances of winning the season.

Returning Players

Hypothesis

If you are a returning player are you more likely to win?
Null Hypothesis: being a returning player has no effect on your chances of winning
Alternative Hypothesis: being a returning player has an effect on your chances of winning

Hypothesis Testing

```r
# If you are a returning player, are you more likely to be voted off?
# find proportion of returning players that won
name.counts <- sort(table(castaways$full_name), decreasing = TRUE)
return_cast <- names(which(name.counts > 1))
return_total <- length(return_cast)
sole_name <- subset(castaways, result == "Sole Survivor")
sole_name <- sole_name$full_name
return_sole <- intersect(return_cast, sole_name)
total_sole <- length(sole_name)
total_return_sole <- length(return_sole)

# sample proportion of returning players
prop_returning_players <- total_return_sole / total_sole

# sample proportion of non-returning players
not_return_sole <- total_sole - total_return_sole
prop_nonreturn_sole <- not_return_sole / total_sole
```

Results:

```
> prop_returning_players
[1] 0.5666667
> prop_nonreturn_sole
[1] 0.4333333
>
```

Returning: 56.66%
Non Returning: 43.33%

After determining the proportions we can generate the z-score and find the p-value to see if we accept or reject our null hypothesis.

```r
# calculate SE
SE <- sqrt(prop_returning_players * (1 - prop_returning_players) / return_total +
           prop_nonreturn_sole * (1 - prop_nonreturn_sole) / return_total)

# calculate z-score
z <- (prop_returning_players - prop_nonreturn_sole) / SE

# calculate p-value (two tailed test)
p_value <- 2 * (1 - pnorm(abs(z)))
```

```
> SE
[1] 0.05343498
> z
[1] 2.495244
> p_value
[1] 0.01258704
```

z-Score = 2.49, p-value = .012, which is less than the alpha .005

Based on our findings we reject our null hypothesis because there is enough evidence to show that if someone is a returning player that their chances of winning Sole Survivor goes up by 13.33%

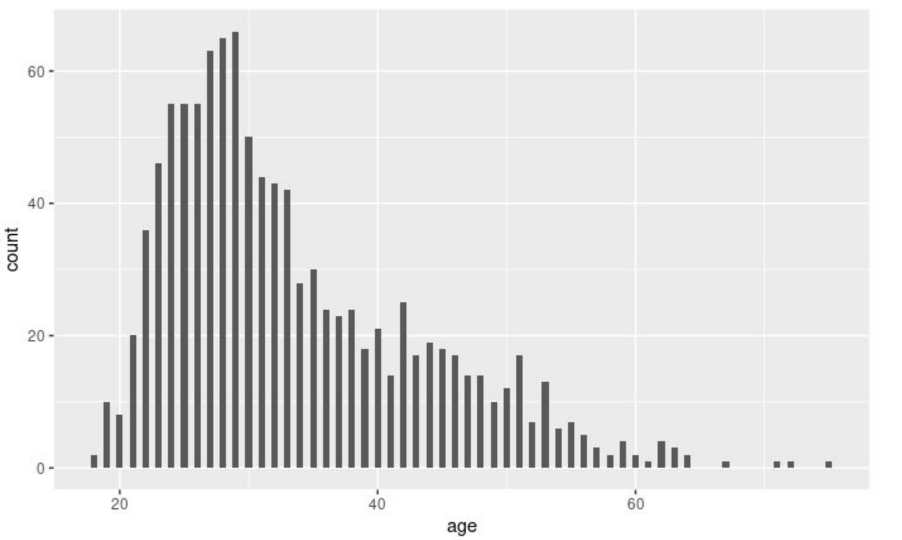
Age

Hypothesis

Does your age affect your chances of winning?
Null Hypothesis: age has no effect on your chances of winning
Alternative Hypothesis: age has an effect on your chances of winning

Histogram and Summary Table: Non winners

```{r}
#age of non winners histogram
ggplot(data = not_winners_age, aes(x = age)) + geom_histogram(binwidth = .5)
```



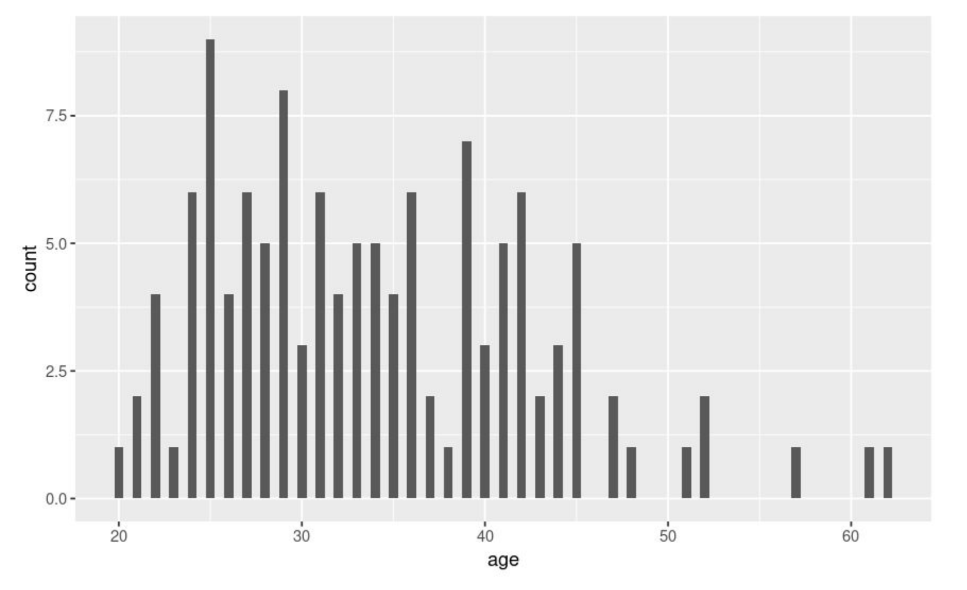

```{r}
summary(not_winners_age$age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  18.00   26.00   31.00   33.41   39.00   75.00       3 
```

Histogram and Summary Table: Winners

```{r}
ggplot(ageWin, aes(x = age)) + geom_histogram(binwidth = .50)

summary(ageWin$age)
```



Min	1st Qu.	Median	Mean	3rd Qu.	Max.
20.00	27.00	33.00	34.07	40.00	62.00

Test the hypothesis:

To test the hypothesis (Does the rate of challenge wins affect your chances of winning?) The following is how the null and alternative hypothesis was formatted:
Null_Hypothesis: Age does not have an impact on chances of winning.
Alternative_Hypothesis: Age does have an impact on chances of winning.

To test this hypothesis, an inference for the difference of 2 means test was used

````
#hypothesis testing
```{r}
pwa = mean(winners$age, trim = 0, na.rm = TRUE)
nw = 60
sdw = sd(winners$age)
```

```{r}
pta = mean(not_winners_age$age, trim = 0, na.rm = TRUE)
nt = 899
sdt = sd(not_winners_age$age %>% na.omit())
```

```{r}
SE_age = sqrt((sdw*sdw)/nw + (sdt*sdt)/nt)
```

```{r}
df = nw - 1 + nt - 1
```

```{r}
t = (pwa - pta)/SE_age
```

```{r}
p_value_age = 2*(pt(t, df))
```
````

````
#Confidence interval
```{r}
t_star = qt(.95 + .05/2, df)
```

```{r}
ME_age = t_star*SE_age
```

```{r}
CI_lower_age = pwa - pta - ME_age
CI_upper_age = pwa - pta + ME_age
```
````

P-value = .4101. A confidence level of 95% was chosen, which means alpha is .05. Because the p-value is higher than alpha, we fail to reject the null hypothesis.

Check for error: (-3.1881, 1.2998). This means that there is a 95% chance that the difference between the two means is between -3.1881 and 1.2998. This confirms our acceptance of the null hypothesis.

Conclusions

Based on our findings our confidence interval is (-3.1881, 1.2998). This means that there is a 95% chance that the difference between the two means is between -3.1881 and 1.2998. This confirms our acceptance of the null hypothesis that age does not have an impact on winning.

Personality Type

Hypothesis

Does your personality type affect your chances of winning?
Null Hypothesis: personality type has no effect on your chances of winning
Alternative Hypothesis: personality type has an effect on your chances of winning

Creating the Data Frame

We first found the number of each personality type in both winners and non winners:

```
```{r}
not_winners_personality <- not_winners_p %>% group_by(personality_type) %>% tally()
colnames(not_winners_personality)[2] ="not_winners"
glimpse(not_winners_personality)
```
```

```
Rows: 17
Columns: 2
$ personality_type <chr> "ENFJ", "ENFP", "ENTJ", "ENTP", "ESFJ", "ESFP", "ESTJ", "ESTP", "INFJ…
$ not_winners      <int> 37, 80, 34, 44, 37, 70, 48, 47, 23, 54, 23, 39, 40, 64, 54, 33, 344
```

```
```{r}
winners_personality <- winners %>% group_by(personality_type) %>% tally()
colnames(winners_personality)[2] ="winners"
glimpse(winners_personality)
```
```

```
Rows: 17
Columns: 2
$ personality_type <chr> "ENFJ", "ENFP", "ENTJ", "ENTP", "ESFJ", "ESFP", "ESTJ", "ESTP", "INFJ…
$ winners          <int> 3, 3, 1, 5, 1, 3, 4, 8, 4, 1, 2, 2, 2, 2, 1, 2, 16
```

Next, we combined these two tables to create a comparison table. We also added a column to compute the expected number of winners, and a z score to compare the expected and the result:

```
```{r}
#combining two columns
all_contestants_personality = data.frame(not_winners_personality, winners_personality$winners)
#rename column
colnames(all_contestants_personality)[3] ="winners"
#filter out NAs
all_contestants_personality <- all_contestants_personality|>
  filter(personality_type != "NA")
#add total row
all_contestants_personality[nrow(all_contestants_personality) + 1,] <- c("total", 555, 44)
#convert values to integers
all_contestants_personality$not_winners <- strtoi(all_contestants_personality$not_winners)
all_contestants_personality$winners <- strtoi(all_contestants_personality$winners)
#create expected column
all_contestants_personality$expected <- (44/555)*(all_contestants_personality$not_winners)
#finding z scores
all_contestants_personality$z <-
  (all_contestants_personality$winners -
all_contestants_personality$expected)/sqrt(all_contestants_personality$expected)

glimpse(all_contestants_personality)
```
```

	personality_type	not_winners	winners	expected	z
1	ENFJ	37	3	2.933333	0.03892495
2	ENFP	80	3	6.342342	-1.32716858
3	ENTJ	34	1	2.695495	-1.03270751
4	ENTP	44	5	3.488288	0.80939924
5	ESFJ	37	1	2.933333	-1.12882347
6	ESFP	70	3	5.549550	-1.08226743
7	ESTJ	48	4	3.805405	0.09975400
8	ESTP	47	8	3.726126	2.21407805
9	INFJ	23	4	1.823423	1.61187065
10	INFP	54	1	4.281081	-1.58577014
11	INTJ	23	2	1.823423	0.13076434
12	INTP	39	2	3.091892	-0.62096553
13	ISFJ	40	2	3.171171	-0.65767379
14	ISFP	64	2	5.073874	-1.36463407
15	ISTJ	54	1	4.281081	-1.58577014
16	ISTP	33	2	2.616216	-0.38097485
17	total	555	44	44.000000	0.00000000

Hypothesis Testing

Does your personality type affect your chances of winning?

I computed a p-value using my z values and a Chi-Square test

```{r}
Chi_square <- sum(all_contestants_personality$z*all_contestants_personality$z)
```

```{r}
p_value_personality <- 1 - pchisq(Chi_square, 15)
```

Conclusion

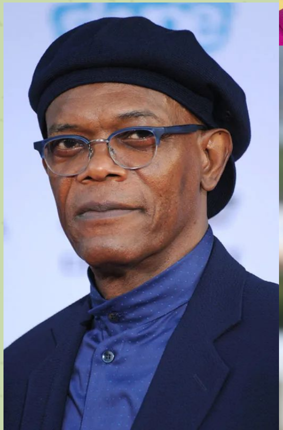
We found a p-value of .1271 which meant there was not enough evidence to reject the null hypothesis.
Therefore, the rate of challenge wins does not impact your chances of winning the season.
However, there was one personality type with a high core meaning these people won more than expected.

Best Personality Type

The personality type with the highest z score of 2.214 is the ESTP or entrepreneur personality type.

This personality is characterized as

- Energetic
- Observant
- Impulsive
- Competitive



Conclusion

Conclusion

Based on our findings we discovered that while having more challenge wins gave these Survivors a leg up, challenge wins, age, and personality types doesn't have a substantial impact on winning Survivor. However, being a retuner player does. We conclude that in the end one's social game has the most impact on being the winner.

Challenge Wins ✗

Returning Player ✓

Age ✗

Personality Type ✗