# DATA557 Homework 3

*Will Wright*

*January 26, 2019*

## Instructions

Submit your solutions in pdf format to the dropbox on the canvas page by 5:00PM, Wednesday January 30. You do not need to include your R code with your solutions for this assignment.

For question 1, you are to work in groups. For the other questions, you may work together to help each other solve problems, but you should do all the work, create your own solutions, and hand in your own work without copying others' work.

### Question 1. (This is Q3 of Exercise 3.)

Suppose that a new experiment is being designed to determine the effect on output of temperatures higher than 100. In particular, the aim of the new experiment is to test the null hypothesis that the mean output is the same for temperature 100 and temperature 120. The researcher would like to have at least 90% power to detect a difference between these conditions in mean output equal to 75. Your job is to determine the sample sizes for each group and to decide which test statistic will be used to test the null hypothesis. Justify your answers.

For this question, you should continue to work with your group on the answer that you developed in class on Jan 23. The group member who did the original posting will receive feedback from me on canvas by Saturday morning and should relay it to the other group members. Work together either in person or electronically to develop your final solution. The group member who did the original posting should include the group's solution in their HW 3 submission, including the names of all group members. The grade assigned for the solution to this question will be applied to the grade for the HW for all group members.

### Question 1 Findings:

In our approach to this question, we first inspected the known experimental parameters to help determine which statistical test to use. In particular, we were interested in what we could learn about the distributions from the two samples of n=30 from the prior experiment for temperatures of 50 and 100.

```
# read file and set seed for reproducibility
pDat <- read.csv("../WEEK03/process.csv")
set.seed(20190127)

# subset and calculate means and sds
data_50 <- pDat$output[which(pDat$temp==50)]
data_100 <- pDat$output[which(pDat$temp==100)]
mean_50 <- mean(data_50)
mean_100 <- mean(data_100)
sd_50 <- sd(data_50)
sd_100 <- sd(data_100)
```

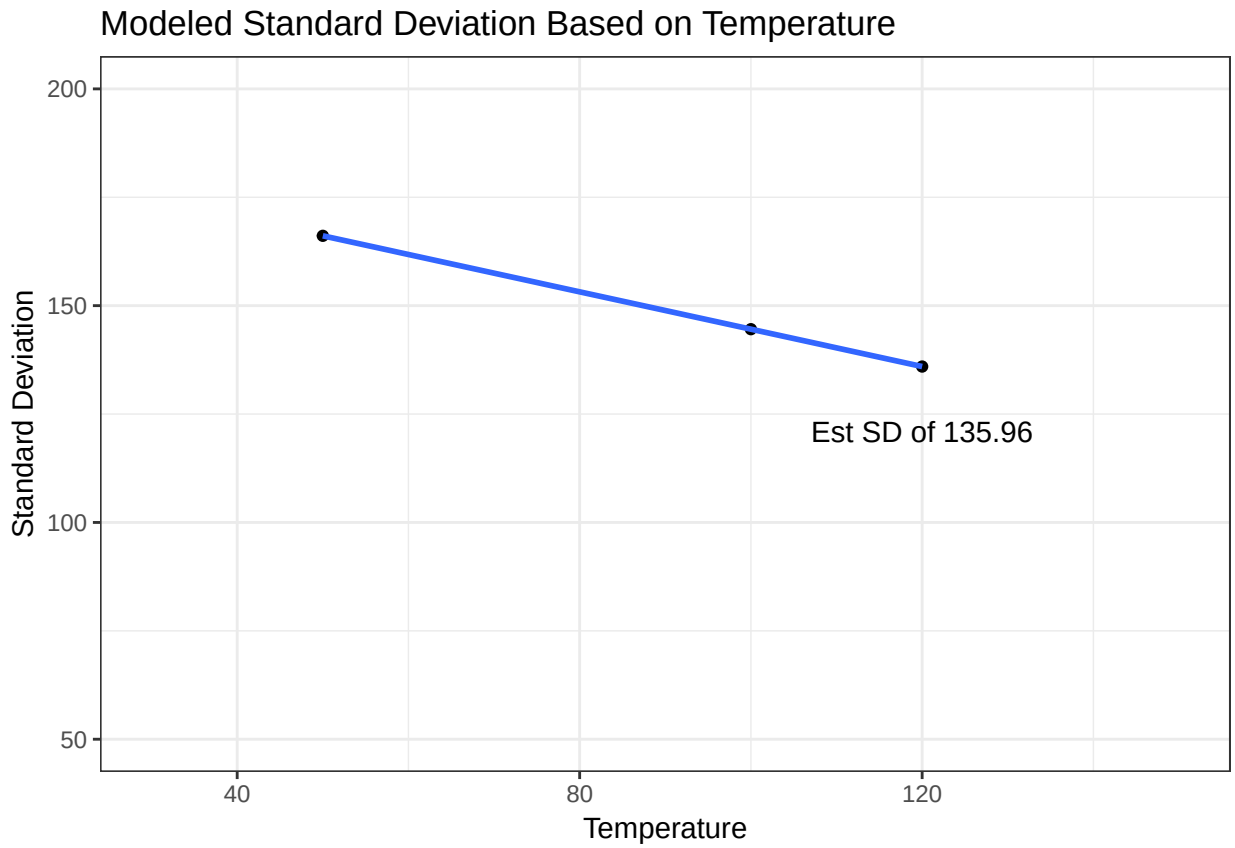
The mean and standard deviation for the sample with a temperature of 50 are 899.82 and 166.11.

The mean and standard deviation for the sample with a temperature of 100 are 1034.14 and 144.57.

Next, in order to estimate the standard deviation for a new sample with a temperature of 120, we used a linear model to extrapolate. Given more data or context about the machine where measurements are taken, we could likely make better assumptions about what standard deviation is reasonable. For instance, if we know that failures happen more frequently above 100 degrees, a linear model would not be appropriate. That said, this approach seems reasonable given the situation.

```
# for temp = 120, use a linear model to estimate based on temp = 50 and temp = 100 data
temps <- c(50,100)
sds <- c(sd_50, sd_100)
fit1 <- lm(sds ~ temps)
sd_120 <- fit1$coefficients[1]+fit1$coefficients[2]*120 # linear model for est sd of temp = 120
sdDat <- data.frame(temps = c(temps, 120), sds = c(sds,sd_120)) #append new datapoint

# plot modeled SD for temp = 120
p <- ggplot(sdDat, aes(x = temps, y = sds))
p + geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw() +
  labs(title = "Modeled Standard Deviation Based on Temperature",
       x = "Temperature",
       y = "Standard Deviation") +
  scale_y_continuous(limits = c(50,200)) +
  scale_x_continuous(limits = c(30,150)) +
  annotate("text", 120, sd_120-15, label = paste0("Est SD of ",round(sd_120,2)))
```

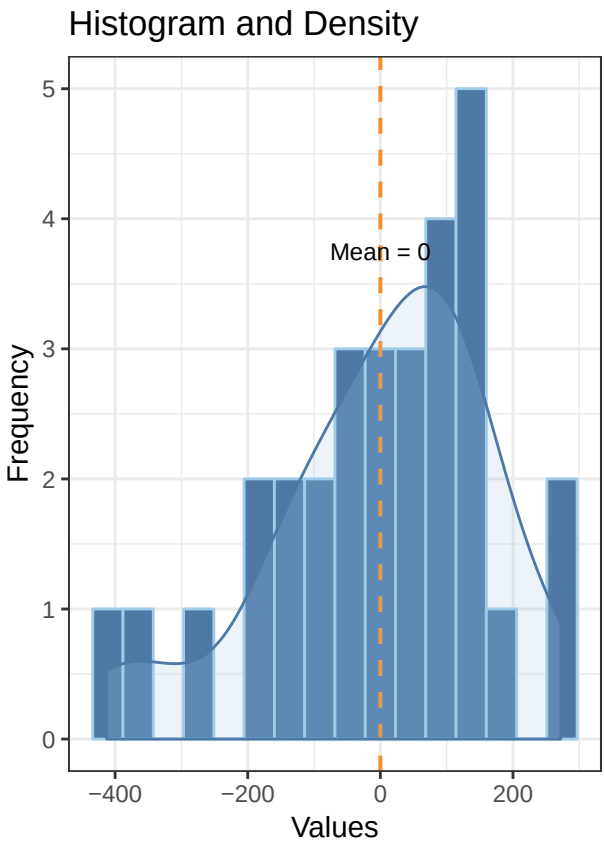


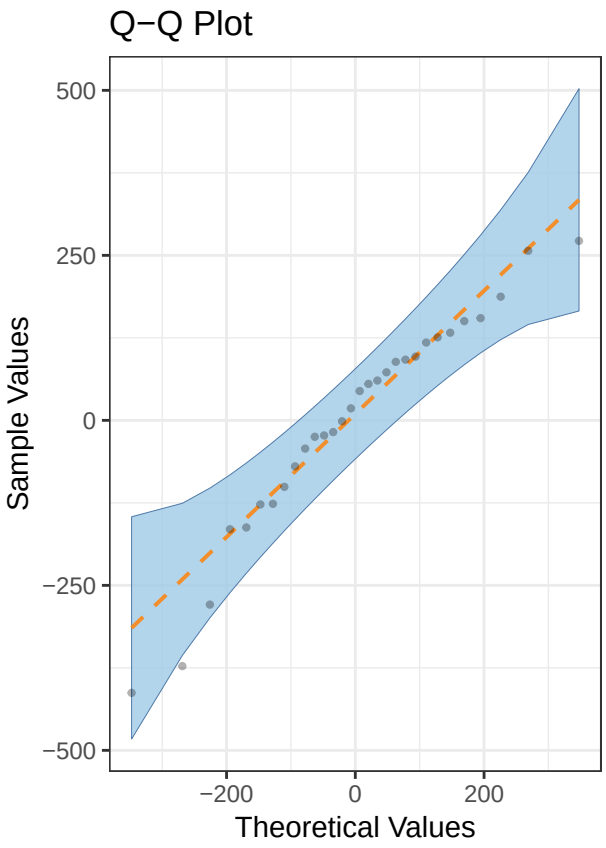

Our model shows an estimated standard deviation of 135.96.

Then, we inspected the distribution of residuals for the temp = 50 and temp = 100 cases to see if normality

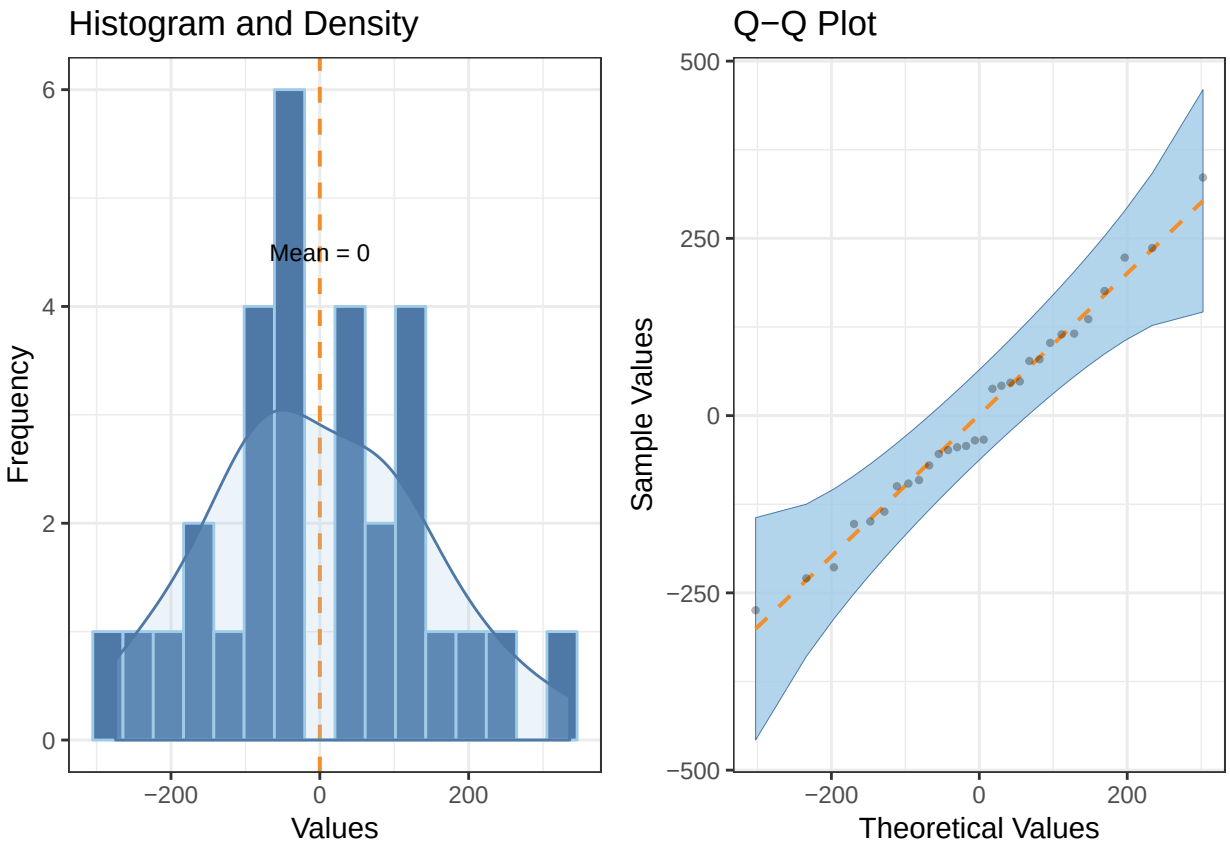
is a reasonable assumption:

```
# inspect residuals
residuals_50 <- data_50 - mean_50
residuals_100 <- data_100 - mean_100
distribution_visualizer(residuals_50)
```




```
distribution_visualizer(residuals_100)
```




Based on these results, normality seems plausible, which satisfies the required assumption to move forward with the following simulations:

To examine the relationship between sample sizes and power, we investigated 3 conditions:

1. Assuming a completely normal distribution with equal variance using a Z-statistic
2. Assuming a t-distribution with equal variance using a T-statistic
3. Assuming a t-distribution with unequal variance using a T-statistic (Welch's T-test)

Below, we simulated 1000 samples in each condition to get a mean power per each sample size from 1 to 100:

```
# calculate input parameters
output_diff <- 75
n_vals <- 1:100
reps=1000

# simulations
powers_z <- rep(NA,length(n_vals))
powers_t <- rep(NA,length(n_vals))
powers_w <- rep(NA,length(n_vals))
for(j in 1:length(n_vals)){
  test_statistic_equalVar <- rep(NA,reps)
  test_statistic_unequalVar <- rep(NA,reps)
  welch_df <- rep(NA,reps)
  for(i in 1:reps){
    temp_100sim <- rnorm(n_vals[j], 75, sd_100)
    temp_120sim_equalVar <- rnorm(n_vals[j], 0, sd_100)
```

```
    temp_120sim_unequalVar <- rnorm(n_vals[j], 0, sd_120)
    se_equalVar <- sqrt(var(temp_100sim)/n_vals[j]+var(temp_120sim_equalVar)/n_vals[j])
    se_unequalVar <- sqrt(var(temp_100sim)/n_vals[j]+var(temp_120sim_unequalVar)/n_vals[j])
    test_statistic_equalVar[i] <- abs(mean(temp_100sim)-mean(temp_120sim_equalVar))/se_equalVar
    test_statistic_unequalVar[i] <- abs(mean(temp_100sim)-mean(temp_120sim_unequalVar))/se_unequalVar
    temp_100sim_sd <- sd(temp_100sim)
    temp_120sim_sd <- sd(temp_120sim_unequalVar)
    welch_df[i] <- (temp_100sim_sd^2/n_vals[j] + temp_120sim_sd^2/n_vals[j])^2/
      (temp_100sim_sd^4/(n_vals[j]^2*(n_vals[j]-1)) + temp_120sim_sd^4/(n_vals[j]^2*(n_vals[j]-1)))
  }
  powers_z[j] <- mean(abs(test_statistic_equalVar)>qnorm(0.975)) # for z-test
  powers_t[j] <- mean(abs(test_statistic_equalVar)>qt(0.975, df = n_vals[j]-2)) # t-test
  welch_df <- mean(welch_df)
  powers_w[j] <- mean(abs(test_statistic_unequalVar)>qt(0.975, df = welch_df)) # welch's t-test
}
```

```
## Warning in qt(0.975, df = n_vals[j] - 2): NaNs produced
```

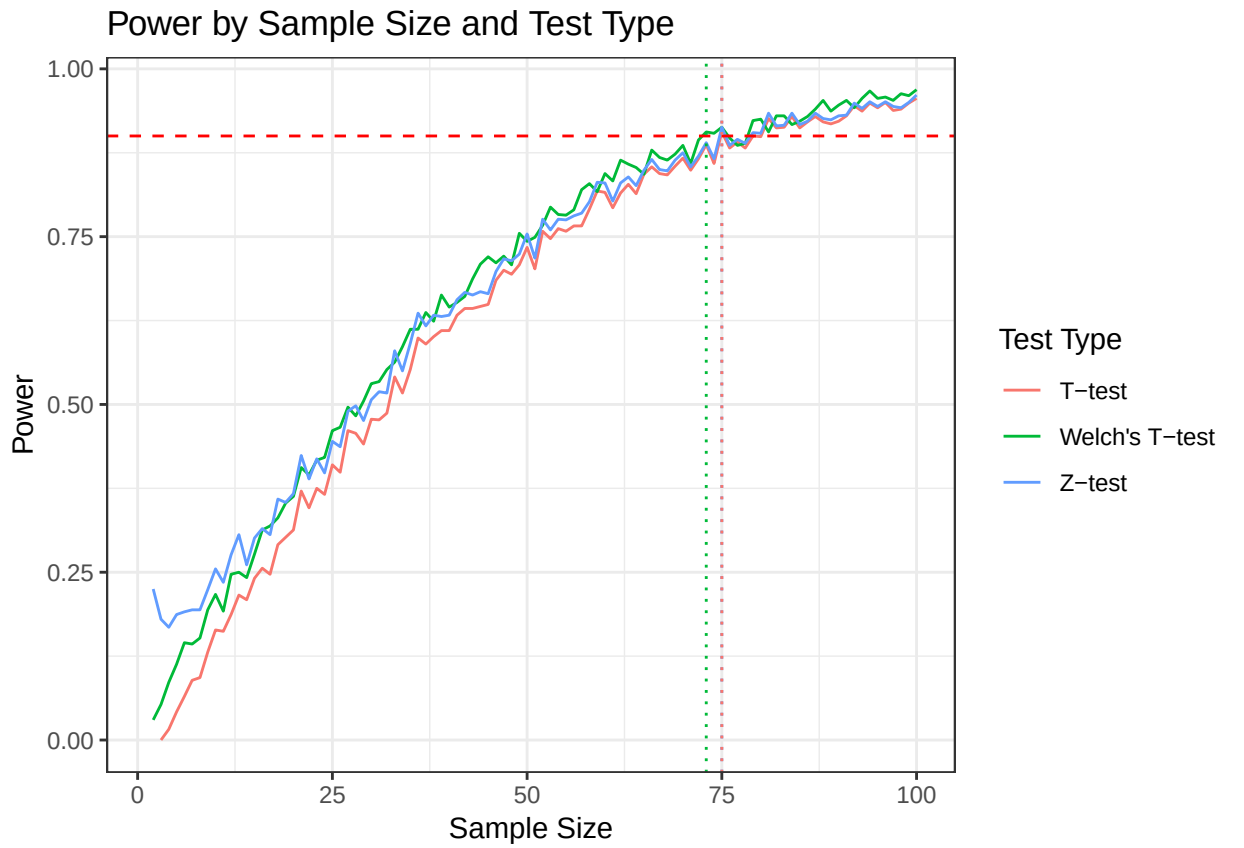
```
## Warning in qt(0.975, df = n_vals[j] - 2): NaNs produced
```

```
results <- data.frame(n = rep(n_vals,3),
                      test_type = c(rep("Z-test", length(n_vals)),
                                    rep("T-test", length(n_vals)),
                                    rep("Welch's T-test", length(n_vals))),
                      power = c(powers_z, powers_t, powers_w))

ideal_n_z <- min(results$n[which(results$power>=0.9 & results$test_type == "Z-test")])
ideal_n_t <- min(results$n[which(results$power>=0.9 & results$test_type == "T-test")])
ideal_n_w <- min(results$n[which(results$power>=0.9 & results$test_type == "Welch's T-test")])

g <- ggplot(results, aes(x = n, y = power, color = test_type))
g + geom_line() +
  scale_x_continuous(limits = c(1,100)) +
  theme_bw() +
  geom_hline(yintercept = 0.9, col = "red", linetype = "dashed") +
  geom_vline(xintercept = ideal_n_z, col = "#619CFF", linetype = 3) +
  geom_vline(xintercept = ideal_n_t, col = "#F8766D", linetype = 3) +
  geom_vline(xintercept = ideal_n_w, col = "#00BA38", linetype = 3) +
  labs(title = "Power by Sample Size and Test Type",
       x = "Sample Size",
       y = "Power") +
  scale_color_discrete(name = "Test Type")
```

```
## Warning: Removed 4 rows containing missing values (geom_path).
```




The results of this power analysis show that the needed sample size is 75, 75, or 73 depending on if we wanted to go with a Z-test, T-test, or Welch's T-test, respectively.

Recommendation: We ruled out the Z-test even before we started since it's the approach with the most assumptions (normality, equal variance, and a sufficiently large sample). Between the T-test and Welch's T-test, we ultimately sided with the sample size of 75 via the results from the T-test simulation with the assumption of equal variance. While it is likely the case that the smaller standard deviation for a temperature of 120 is true given our linear model and we could therefore use a Welch's T-test, we'd prefer to take a more conservative approach and assume the standard deviation is equal to the higher standard deviation at temp = 100.

## Question 2

Data set: 'fev.csv'

The data come from a study to examine, among other things, the association between smoking and lung function as measured by forced expiratory volume (FEV) in children. Variables in the data set are listed below.

id: ID number age: Age (yrs) fev: FEV (liters) ht: Height (inches) male: Sex (0=female, 1=male) smoke: Smoking Status (0=non-current smoker, 1=current smoker)

```
fevDat <- read.csv("../WEEK03/fev.csv")
smokerDat <- fevDat[which(fevDat$smoke==1),]
nonSmokerDat <- fevDat[which(fevDat$smoke==0),]
```

2.1. It is desired to conduct a test of the null hypothesis that mean FEV is the same in smokers and non-smokers. Conduct an appropriate simulation study to determine an appropriate hypothesis test procedure to use (consider only the 3 methods discussed in class). Provide a description of your simuation study in words and provide a summary of the results of your simulation study.

$H_0 : \mu_A = \mu_B$ where A are smokers and B are non-smokers $H_1 : \mu_A \neq \mu_B$

In order to determine an appropriate hypothesis test procedure, we must consider how large the sample size is,

```
# N<-5000
# set.seed(20190126)
# n1<-30
# n2<-15
# n<-n1+n2
# sim_data<-data.frame(temp=c(rep(50,n1),rep(100,n2)),output=rep(NA,n))
# sim_result<- data.frame(p.equal.var=rep(NA,N),p.Welch=rep(NA,N),p.Z=rep(NA,N))
# for(i in 1:N){
#     sim_data$output<-rnorm(n,mean=1000,sd=c(rep(80,n1),rep(20,n2)))
#     result<-with(sim_data,t.test(output[temp==50],output[temp==100],var.equal=T))
#     sim_result[i,1]<-result$p.value
#     result<-with(sim_data,t.test(output[temp==50],output[temp==100],var.equal=F))
#     sim_result[i,2]<-result$p.value
#     Z<-result$statistic
#     sim_result[i,3]<-2*(1-pnorm(Z))
# }
# apply(sim_result<0.05,2,mean)
```

2.2. Using the method chosen in Q2.1, carry out the test of the null hypothesis. Report your results along with a confidence interval.

2.3. Conduct a test of the null hypothesis that mean FEV is the same in smokers and non-smokers, among children 10 years old or older. Compare your results to the previous results for all children and provide an interpretation of any differences in results. (Note: Children younger than 10 would be unlikely to smoke and had been included in the study for other purposes besides the analysis of the association between smoking and FEV.)

## Question 3

Read the paper "Why most published research findings are false." (see Week 3 folder). Do you agree with the overall conclusion of the paper that most published research findings are false? Provide your answer with your rationale in a paragraph of 200 words or less. (No simulation studies required here!)

Based purely on this paper, I'm fine with accepting statement that most published research findings are false. That said, I disagree with the general sentiment it provokes-that we can't trust scientific papers in general. My understanding is that, by volume, most research is victim to the 6 corollaries he defines (i.e. small studies,

small effect sizes, flexible experimental design etc.). As evidence of this, you can find a scientific paper to cite as evidence for nearly any crazy idea (including that vaccines don't work or that human-caused climate change is a myth). So it's true that if I had to draw a random research finding from a hat, I'd bet against it being true. That said, research that has been reprodicible and aggregated into meta-analyses with strong effect sizes are a subset of all research that I believe deserves our confidence. Ioannidis touches on this by stating that "Better powered evidence, e.g., large studies or low-bias meta-analyses, may help..." but he goes on to cast doubt regardless.