The p-Value You Cannot BUY

By

Eugene **Demidenko**

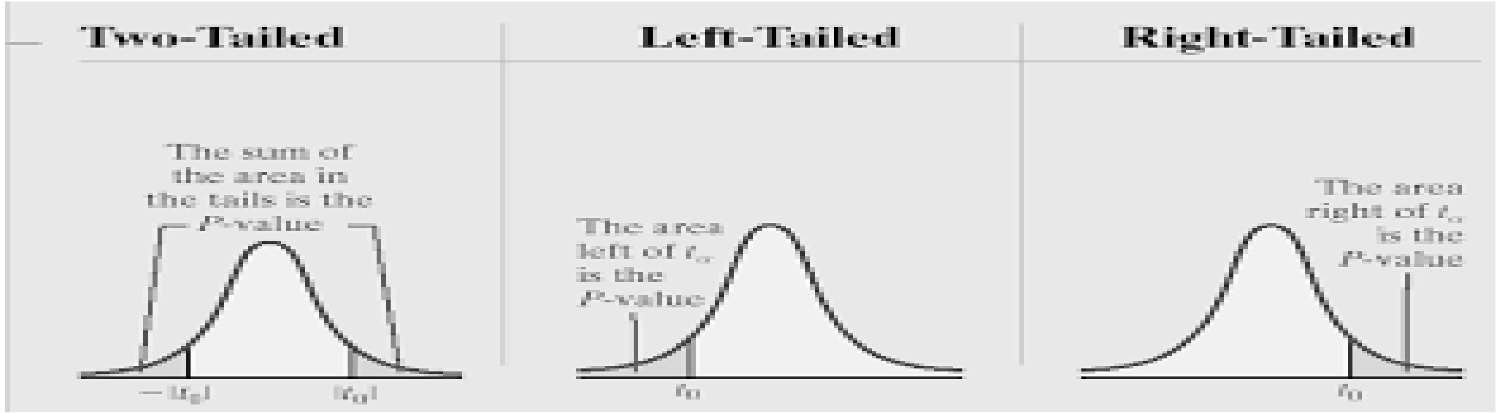
- presented by Diganta (BS1718)

Quick recap of Hypothesis Testing-

- Let $X_1, X_2, X_3, \dots, X_n \sim N(\mu_1, \sigma^2)$ and $Y_1, Y_2, Y_3, \dots, Y_n \sim N(\mu_2, \sigma^2)$ where μ_1, μ_2, σ^2 are all unknown. If we want to test the hypothesis $H_0 : \mu_1 = \mu_2$ against the alternate hypothesis $H_1 : \mu_1 > \mu_2$.
- Then we will take the Test Statistic as : $t = \frac{\left(\frac{X-\overline{Y}}{\sqrt{n}}\right)^{\overline{}}}{s\sqrt{2}} \sim t_{2n-2}$
- We will also have $s^2 = \frac{1}{2n-2}\left(\Sigma_{i=1}^{n}(X_i - \overline{X})^2 + \Sigma_{i=1}^{n}(Y_i - \overline{Y})^2\right)$
- We reject the null hypothesis when p-value is less than 0.05 .

Quick recap on p-Value -

- p-Value is the probability of obtaining results as extreme as the observed results of a hypothesis test assuming the null hypothesis is correct. The exact measure of extremeness depends on how it is tested.
- There are 3 cases-



Two-Tailed
Left-Tailed
Right-Tailed
The sum of the area in the tails is the P-value
$-|t_0|$
$|t_0|$
The area left of t_o is the P-value
t_0
The area right of t_o is the P-value
t_0

- p-value for a Two-Sided test- Probability of getting a statistic <-|t_0| or >|t_0|.
- p-value for a Left-Tailed test- Probability of getting a statistic < t_0.
- p-value for a Right-Tailed test- Probability of getting a statistic > t_0.

An interesting observation-

- Suppose we want to check the effectiveness of an anti-obesity drug that's newly launched in the market.
- We randomly select n obese people and put them in the treatment group and we take another n randomly selected obese people and put them in control group. We continue this trial for a year.
- Problem-H_0 : The drug does not have any effect. vs H_1 : The drug works.
- X_1 , X_2 , … , X_n is the treatment group and Y_1 , Y_2 , … , Y_n is the control group.
- : $t = \frac{\left(\frac{\overline{X}-\overline{Y}}{\sqrt{n}}\right)}{s\sqrt{2}}$ ~ t_{2n-2} where s is the estimate of standard deviation.

An interesting observation-

- From the trial we have the data

$$\bar{X} = 250.0 \text{ lbs}$$

$$\bar{Y} = 249.0 \text{ lbs}$$

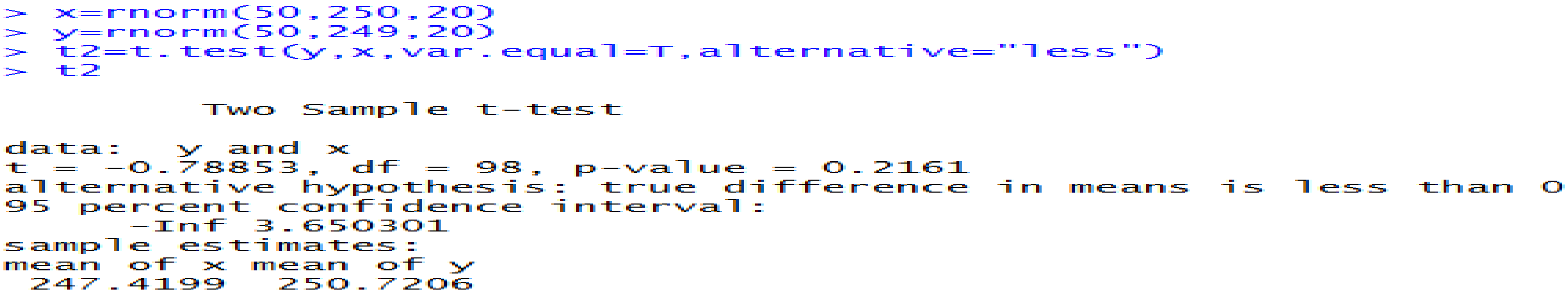
$$\text{Standard Deviation} = s = 20 \text{ lbs}$$

- From this it is pretty clear that the drug does not have any significant effect on weight loss since 1lbs is not significant for someone weighing 250lbs.
- We simulate the following data for, N=50

  N=10,000

Result of the simulation-

- For N=50 , H_0 is not rejected, that means the drug is not effective.

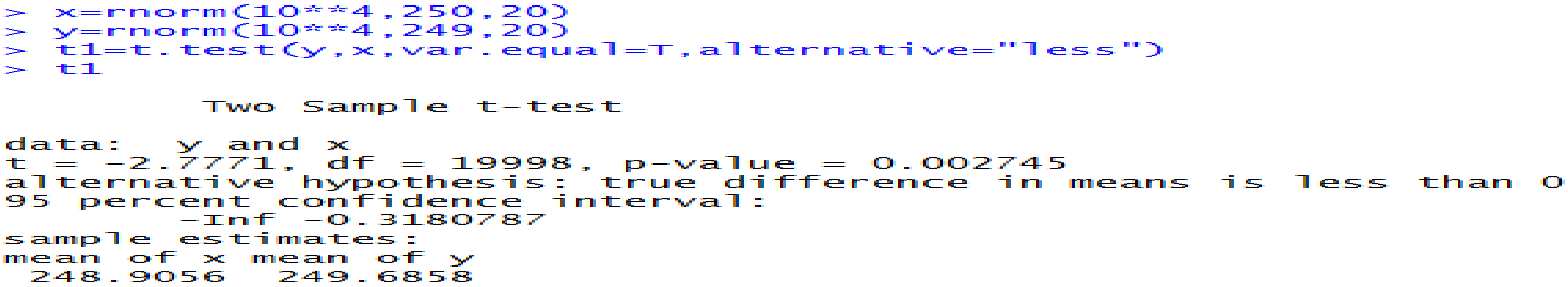


```
> x=rnorm(50,250,20)
> y=rnorm(50,249,20)
> t2=t.test(y,x,var.equal=T,alternative="less")
> t2

        Two Sample t-test

data:  y and x
t = -0.78853, df = 98, p-value = 0.2161
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 3.650301
sample estimates:
mean of x mean of y 
 247.4199  250.7206 
```

- For N=10,000 , H_0 is rejected, meaning the drug is effective.



```
> x=rnorm(10**4,250,20)
> y=rnorm(10**4,249,20)
> t1=t.test(y,x,var.equal=T,alternative="less")
> t1

        Two Sample t-test

data:  y and x
t = -2.7771, df = 19998, p-value = 0.002745
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.3180787
sample estimates:
mean of x mean of y 
 248.9056  249.6858 
```

Another interesting observation-

- Suppose a cancer researcher has developed a new drug and tested its effect on mice.
- He took X_1 , X_2 , … , X_n as the treatment group and Y_1 , Y_2 , … , Y_n as the control group
- From the data we have

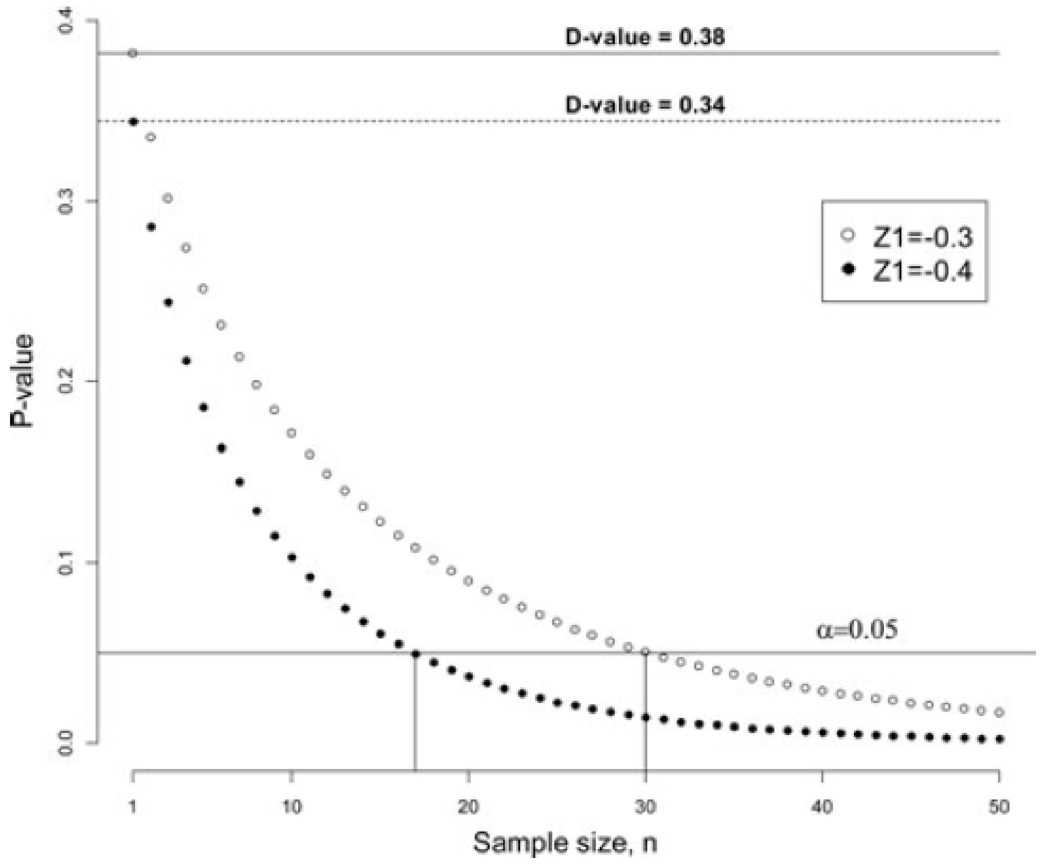
$\overline{X}$ = 15 days , $\overline{Y}$ = 10 days , Standard Deviation = 6 days

EXPERIMENT 1	**EXPERIMENT 2**
n=7	n=14
p-value=0.06	p-value=0.014
The drug is not effective	The drug is effective

Why is this happening?

- The sample size n has a direct effect on the p-value, p-value decreases with increasing sample size.
- Even if the difference between two groups is very small with large enough sample size the absolute value of the test statistic can be made as large as we want and so, the p-value can be made as small as we need.
- This results in a huge number of statistically significant experiments which have no practical significance.

Why is this happening?(Graph)



Reasons for decreasing p-Value -

- The distribution of test statistic becomes leptokurtic (sharp) i.e. the central tendency increases. So, p value decreases with n. That means whenever there is a slight departure from the null hypothesis, with large enough sample size we will eventually reject H_0.
- Our conventional testing procedures which use p-value don’t care how small this departure from H_0 is.
- When n is large it will reject H_0 based on even very small variation of the test statistic.

A new approach –

- Root of the problem with p-value stems form the fact that we are comparing the means of 2 groups. Many researchers from biological fields have challenged mean comparison. The mean may not be the representative of the distribution.
- The idea is when applying a drug to an individual we should estimate the probability that the individual will benefit from the drug and not what the effect on the population is. The question is how does the drug affect a particular person.
- We should be able to provide the risk benefit analysis of treatment based on an individual not a group. We do not want to treat a group we want to treat an individual.

Estimating effect on an individual-

- Consider the obesity experiment for the observation slide. We want to answer - “What is the probability that randomly picked individual will benefit from the drug?”
- To assess the risk and benefit of the treatment, pick at random an individual from the placebo group with weight X and an individual from the drug group with weight Y; where two groups are represented by densities of the weight. The benefit of taking the drug is the probability that Y < X, or symbolically

  The benefit of the drug is $b = P\,(Y < X)$

  The risk of the drug is $\delta = P\,(Y > X)$
- We interpret δ as the probability that a randomly chosen person from the treatment group will be heavier than a randomly chosen person from the placebo group.

An approximation –

- The test statistic is $t = \frac{\left(\frac{X-Y}{\sqrt{n}}\right)}{s\sqrt{2}}$ ~t_{2n-2} ,when n is large we can approximate t by z which follows a Normal(0,1) distribution.

  (∵ Var($\overline{X}$ – $\overline{Y}$)=$\frac{2}{n}\sigma^2$ and s^2 is an unbiased estimator of σ^2).
- So from now on we will assume t follows Normal(0,1). We will define our alternate p-value assuming this, but it can also be extended to *t*-distribution.

D value-

- Risk of the drug is defined as $\delta = P\,(Y > X)$,

  where $Y \sim N\left(\mu_y, \sigma^2\right)$, $X \sim N(\mu_X, \sigma^2)$

  $then,\quad Y - X \sim N(\mu_y - \mu_x, 2\sigma^2)$

  Now,$\delta = P(Y - X > 0)$ $= P\left(Z > -\frac{\mu_y - \mu_x}{\sqrt{2}\sigma}\right)$

  $$= P\left(-Z < \frac{\mu_y - \mu_x}{\sqrt{2}\sigma}\right)$$

  $$= \quad \Phi\left(\frac{\mu_y - \mu_x}{\sqrt{2}\sigma}\right)$$

- D value = d = $\Phi\left(\frac{\bar{Y} - \bar{X}}{\sqrt{2}s}\right)$

P-value vs D-value -

- If {Y_i} and {X_j } are two random samples, conditional on the data, we state,

$$P - value = P(\bar{Y} > \bar{X})\ ;\ D - value = P(Y > X)$$

D value focuses on the effect on an individual not a group.

So, it can be used to prescribe personalized medicines for individual patients.

- Example: Let D-Value = 0.4 And the probability that the medicine works for an old patient is 1 out of 3. Then the actual probability that the medicine would work is 0.2 So, the probability of the patient actually benefitting from the medicine drops from $\frac{1}{3}$ to $\frac{1}{5}$. $\left(\because 0.6 * \frac{1}{3} \ = 0.2\right)$

P value cant give us these type of probabilistic interpretations.

Probabilistic interpretation of D-value

- The D- value is easy to interpret: For example, a widely used effect size of 0.5 means that the proportion of treated patients who do not improve will be roughly 30% and the proportion who do improve will be 70%.

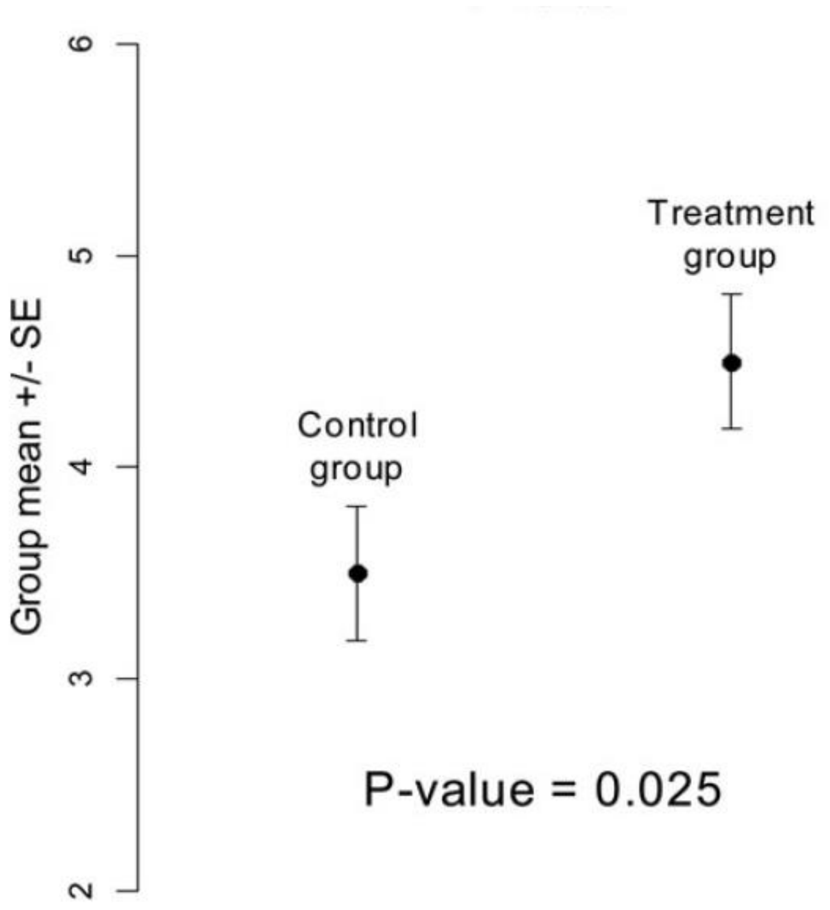
  (D-value = $\Phi(-0.5) = 0.3$).

- Expressing the treatment effect using probability could be important for probabilistic comparison when effect size is not available. For example, consider a typical situation when weighing the pros and cons of a new drug with the D-value = 0.3 for an elderly patient whose chance of survival within 5 years, even if the drug would help, is 1 out of 5. Then the actual benefit of the new drug will be only $0.7 \times 0.2 = 0.14$. Certainly, doctors consider the age of the patient before prescribing a new drug, but the D-value facilitates quantitative assessment of the benefits on the probability scale.

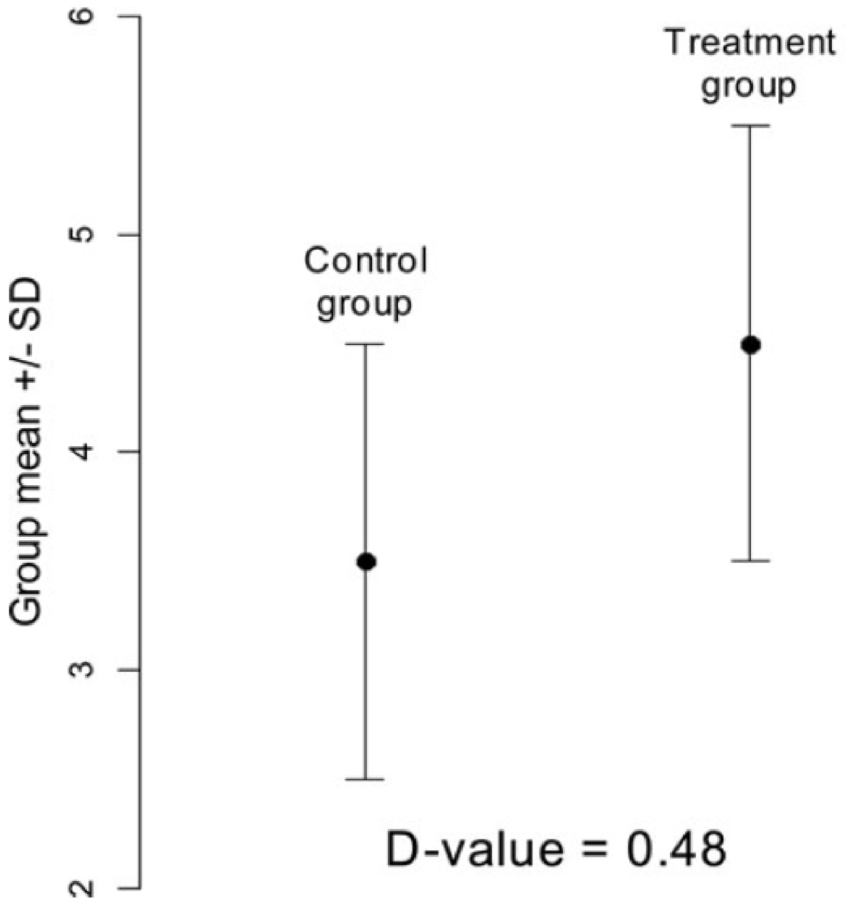
Improved S.E bounds -

- The root of the problem with the p-value is that it compares averages. Since the standard error (SE) of the mean is SD/√n, SE may be as small as desired if the sample size, n, is large enough. When the data are presented graphically as means ± SE, the individual variation is reduced by a factor of the square root of n.
- We should use SD not SE. $\left(\because SE(\bar{X}) = \frac{SD(X)}{\sqrt{n}} \to 0 \; as \; n \to \infty\right)$
- Showing SE error bars silently assumes application of the p- value for group comparison, showing SD error bars assumes the D-value for individual comparison. Since we advocate for the D-value, we use the SD, not SE, error bars. In short, the p- and D-values are computed in the same way but the former uses SE and the latter uses SD.

Improved S.E bounds –



In this figure , SE bounds create illusion of satisfactory separation between the 2 groups with p value 0.025



In this one, we can clearly see the 2 groups are not that different. D value = 0.48

Linear Regression -

$Y_i = \alpha + \beta X_i + \epsilon_i$

Y_i *is dependent variable,*

X_i *is the associated predictor* where $\epsilon_i \sim iid\ N(0, \sigma^2)$

Test: $\mathrm{H}_0\!: \ \beta = 0$ Vs $\mathrm{H}_1\!: \ \beta \neq 0$

- Test statistic: $t = \frac{b}{s} \sim t_{n-2}$
- $s = \frac{\sqrt{\frac{\Sigma\left(Y_i - \hat{Y}\right)^2}{n-2}}}{\sqrt{\Sigma(X_i - \bar{X})^2}} = \sqrt{\frac{1}{n-2}\ \frac{\Sigma(Y_i - \hat{Y})^2}{\Sigma(X_i - \bar{X})^2}}$

Linear Regression –

- The arguments against the traditional p-value in a two-group comparison can be generalized to the linear regression model :

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

where Y_i is the dependent variable, X_i is the associated factor/predictor of the i-th subject or measurement, and ε_i is the normally distributed random error term. The p-value for the slope, β, can be made arbitrarily small by increasing the number of observations, n.

- Since the difference in the means is estimated as b, the slope of the least-square regression, the D-value, is defined as $\Phi\left(-\frac{|b|}{s\sqrt{n}}\right)$ where s is the standard error of the slope from the regression estimation.

Linear Regression –

- P value = $\Phi(-\frac{|b|}{s})$

- D-value = $\Phi\left(-\frac{|b|}{s\sqrt{n}}\right)$

- Just like previous examples $P\ value \rightarrow 0\ as\ n \rightarrow \infty$

- However D-value is independent of sample size.

Justification-

- Consider two populations of y corresponding to x and $x+1$
- The difference in y- means = b.1 = b $(\because \, slope \, is \, b)$
- So by our previous definition,

$$\text{D value} = \text{d} = \Phi\left(-\frac{|b|}{s\sqrt{n}}\right)$$

An example on Regression

- Y : travel time to the nearest cancer centre for *n* = 47,383 breast cancer patients
- Predictors:
  1. Age (numerical)
  2. Stage of cancer (categorical with 4 categories)
  3. Surgery (dichotomous variable) (whether patient has undergone surgery or not)

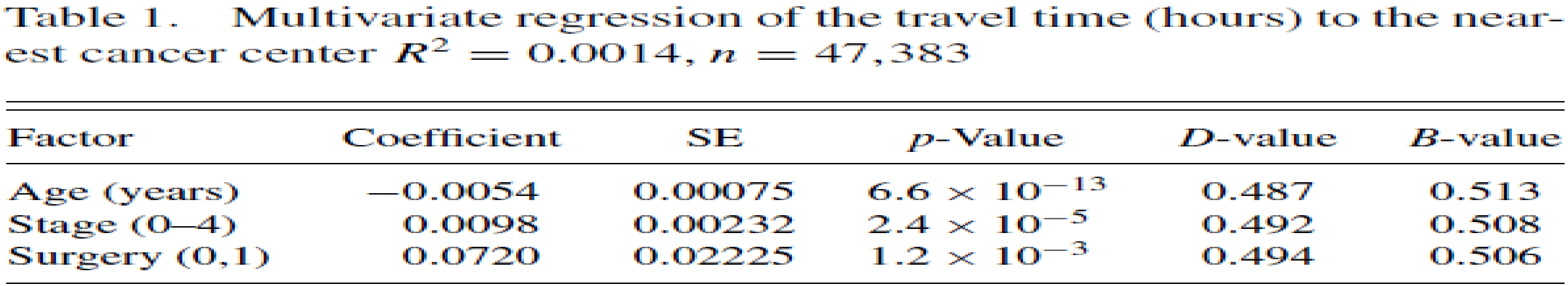


Table 1. Multivariate regression of the travel time (hours) to the nearest cancer center $R^2 = 0.0014$, $n = 47{,}383$

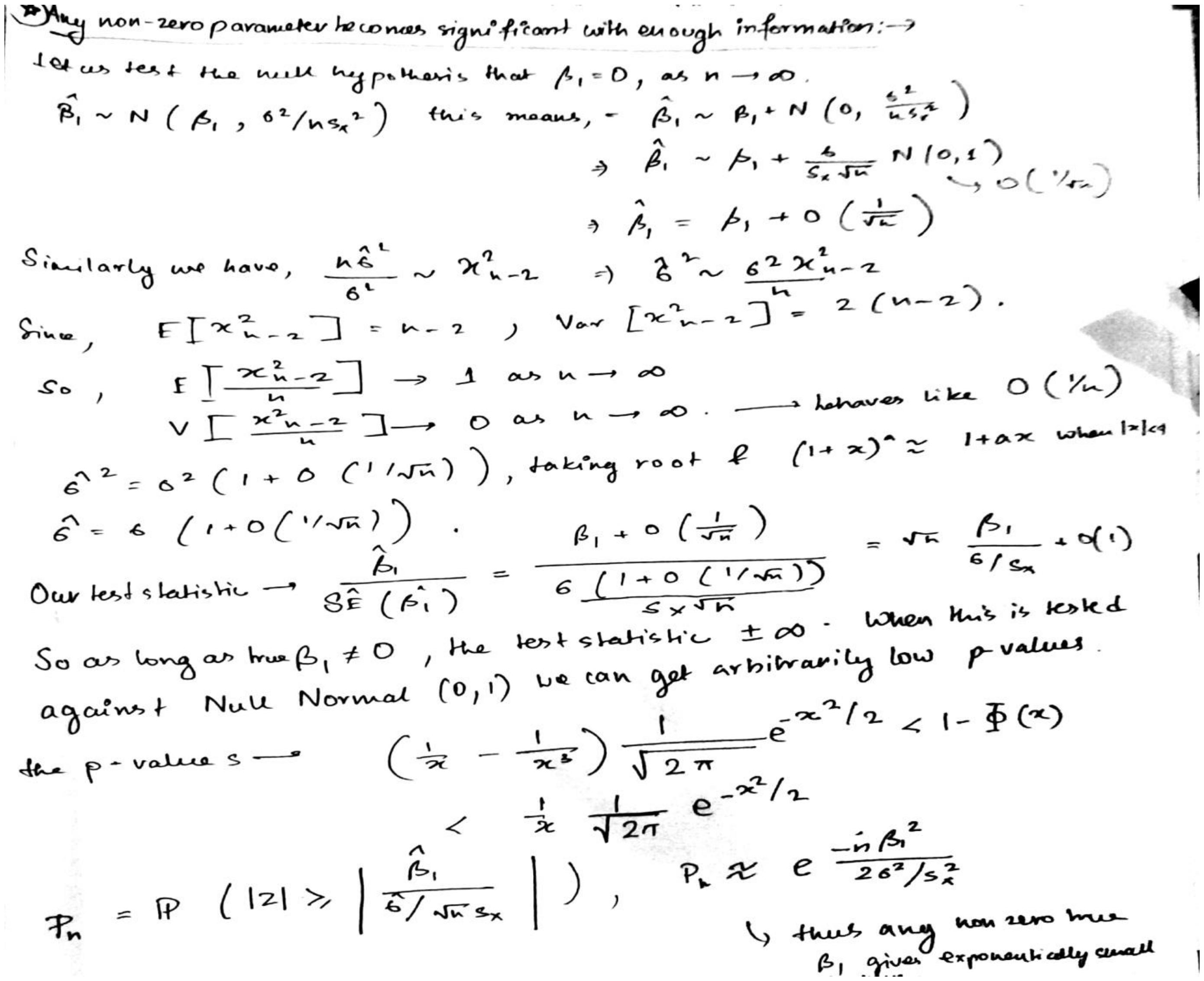
Factor	Coefficient	SE	p-Value	D-value	B-value
Age (years)	−0.0054	0.00075	6.6×10^{-13}	0.487	0.513
Stage (0–4)	0.0098	0.00232	2.4×10^{-5}	0.492	0.508
Surgery (0,1)	0.0720	0.02225	1.2×10^{-3}	0.494	0.506

Observations from the example-

- The *p*-values for all three factors are very small (the factors are statistically significant)
- Yet $R^2 = 0.0014$ i.e. only 0.15% of travel time variation can be explained by these three factors.
- How is it possible?

The answer is large *n*. Paradoxically, the regression may explain almost nothing and yet all predictors may be statistically significant.

Why is this happening?(mathematically)



*) Any non-zero parameter becomes significant with enough information:→

Let us test the null hypothesis that $\beta_1 = 0$, as $n \to \infty$.

$\hat{\beta}_1 \sim N(\beta_1, \sigma^2/n s_x^2)$ this means, - $\hat{\beta}_1 \sim \beta_1 + N(0, \frac{\sigma^2}{n s_x^2})$

$$\Rightarrow \hat{\beta}_1 \sim \beta_1 + \underbrace{\frac{\sigma}{s_x\sqrt{n}} N(0,1)}_{O(1/\sqrt{n})}$$

$$\Rightarrow \hat{\beta}_1 = \beta_1 + O\left(\frac{1}{\sqrt{n}}\right)$$

Similarly we have, $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2} \Rightarrow \hat{\sigma}^2 \sim \frac{\sigma^2 \chi^2_{n-2}}{n}$

Since, $E[\chi^2_{n-2}] = n-2$, $Var[\chi^2_{n-2}] = 2(n-2)$.

So, $E\left[\frac{\chi^2_{n-2}}{n}\right] \to 1$ as $n \to \infty$

$V\left[\frac{\chi^2_{n-2}}{n}\right] \to 0$ as $n \to \infty$. ⟶ behaves like $O(1/n)$

$\hat{\sigma}^2 = \sigma^2(1 + O(1/\sqrt{n}))$, taking root & $(1+x)^a \simeq 1 + ax$ when $|x|<1$

$\hat{\sigma} = \sigma(1 + O(1/\sqrt{n}))$.

Our test statistic → $\frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)} = \frac{\beta_1 + O(\frac{1}{\sqrt{n}})}{\sigma\frac{(1 + O(1/\sqrt{n}))}{s_x\sqrt{n}}} = \sqrt{n}\frac{\beta_1}{\sigma/s_x} + O(1)$

So as long as true $\beta_1 \neq 0$, the test statistic $\pm\infty$. When this is tested against Null Normal $(0,1)$ we can get arbitrarily low p-values.

the p-values → $\left(\frac{1}{x} - \frac{1}{x^3}\right)\frac{1}{\sqrt{2\pi}}e^{-x^2/2} < 1 - \Phi(x)$

$$< \frac{1}{x}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$

$P_n = \mathbb{P}\left(|Z| \geq \left|\frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{n}s_x}\right|\right)$, $P_n \approx e^{\frac{-n\beta_1^2}{2\sigma^2/s_x^2}}$

↳ thus any non zero true β_1 gives exponentially small

We have, $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/ns_X^2)$

$$
\begin{aligned}
\text{So, } \hat{\beta}_1 &\sim \beta_1 + N(0, \sigma^2/ns_X^2) \\
&= \beta_1 + \frac{\sigma}{s_X\sqrt{n}} N(0,1) \\
&= \beta_1 + O(1/\sqrt{n})
\end{aligned}
$$

Also on the other hand we have,

$$
\begin{aligned}
n\hat{\sigma}^2 &\sim \sigma^2 \chi^2_{n-2} \\
\hat{\sigma}^2 &\sim \sigma^2 \frac{\chi^2_{n-2}}{n}
\end{aligned}
$$

Since $\mathbb{E}\left[\chi^2_{n-2}\right] = n-2$ and $\mathrm{Var}\left[\chi^2_{n-2}\right] = 2(n-2)$,

$$
\begin{aligned}
\mathbb{E}\left[\frac{\chi^2_{n-2}}{n}\right] &= \frac{n-2}{n} \to 1 \\
\mathrm{Var}\left[\frac{\chi^2_{n-2}}{n}\right] &= \frac{2(n-2)}{n^2} \to 0
\end{aligned}
$$

Variance is $\mathrm{Var}\left[\frac{\chi^2_{n-2}}{n}\right] = O(1/n).$

$$\hat{\sigma}^2 = \sigma^2\left(1 + O(1/\sqrt{n})\right)$$

Taking square root and applying Bernoulli,s theorem to get,

$$\hat{\sigma} = \sigma\left(1 + O(1/\sqrt{n})\right)$$

Putting all of them together in our test statistic for the hypothesis $\beta_1 = 0$,

$$
\begin{aligned}
\frac{\hat{\beta}_1}{\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]} &= \frac{\beta_1 + O(1/\sqrt{n})}{\frac{\sigma\left(1+O(1/\sqrt{n})\right)}{s_X\sqrt{n}}} \\
&= \sqrt{n}\frac{\beta_1 + O(1/\sqrt{n})}{(\sigma/s_X)\left(1+O(1/\sqrt{n})\right)} \\
&= \sqrt{n}\frac{\beta_1}{\sigma/s_X}\left(1+O(1/\sqrt{n})\right) \\
&= \sqrt{n}\frac{\beta_1}{\sigma/s_X} + O(1)
\end{aligned}
$$

We are saying something likee " this coefficient must really be important because we can measure it really precisely.

$$P_n = \mathbb{P}\left(|Z| \geq \left| \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{n}\, s_x} \right| \right)$$

$$= 2\,\mathbb{P}\left(Z \geq \left| \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{n}\, S_x} \right| \right)$$

$$\leq \frac{2}{\sqrt{2\pi}} \frac{e^{-1/2 \frac{\beta_1^2}{\sigma^2/n s_x^2}}}{\left| \frac{\hat{\beta}_1}{\hat{\sigma} / \sqrt{n}\, s_x} \right|}$$

Using Right hand inequality.

Taking log on both sides, and divide by n,

$$\frac{1}{n} \log P_n \leq \frac{1}{n} \log \frac{2}{\sqrt{2\pi}} - \frac{1}{n} \log \left| \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{n}\, s_x} \right| - \frac{1}{2n} \frac{\beta_1^2}{\sigma^2/n s_x^2}$$

$$= \frac{\log \sqrt{2}\,\pi}{n} + \frac{\log \left| \frac{\hat{\beta}_1}{\hat{\sigma}/S_x} \right|}{n} - \frac{\log n}{2n} - \frac{\hat{\beta}_1^2}{2\hat{\sigma}^2 / S_x^2}$$

taking limit as $n \to \infty$

$$\lim_{n\to\infty} \tfrac{1}{n} \log P_n \leq \lim_n \cancelto{0}{\frac{\log\sqrt{2\pi}}{n}} + \lim_n \cancelto{0}{\frac{\log \hat{\beta}/\hat{\sigma}/S_x}{n}} - \lim_n \cancelto{0}{\frac{\log n}{2n}}$$

$$- \lim_n \frac{\hat{\beta}_1^2}{2\hat{\sigma}^2/S_x^2} \qquad \left[\begin{array}{l} \hat{\beta}_1/(\hat{\sigma}/s_x) \to \beta_1/(\sigma/s_x) \\ \& \ n^{-1} \log n \to 0 \end{array} \right]$$

So, $$\lim_{n\to\infty} \frac{1}{n} \log P_n \leq \frac{-\beta_1^2}{2\sigma^2/s_x^2}.$$

Only using the upper bound we got, $\lim\limits_{n\to\infty} \frac{1}{n} \log P_n \leq \frac{-\beta_1^2}{2\sigma^2/s_x^2}$

Using lower bound we get, $\lim\limits_{n\to\infty} \frac{1}{n} \log P_n \geq \frac{-\beta_1^2}{2\sigma^2/s_x^2}$

So, putting them together we get,

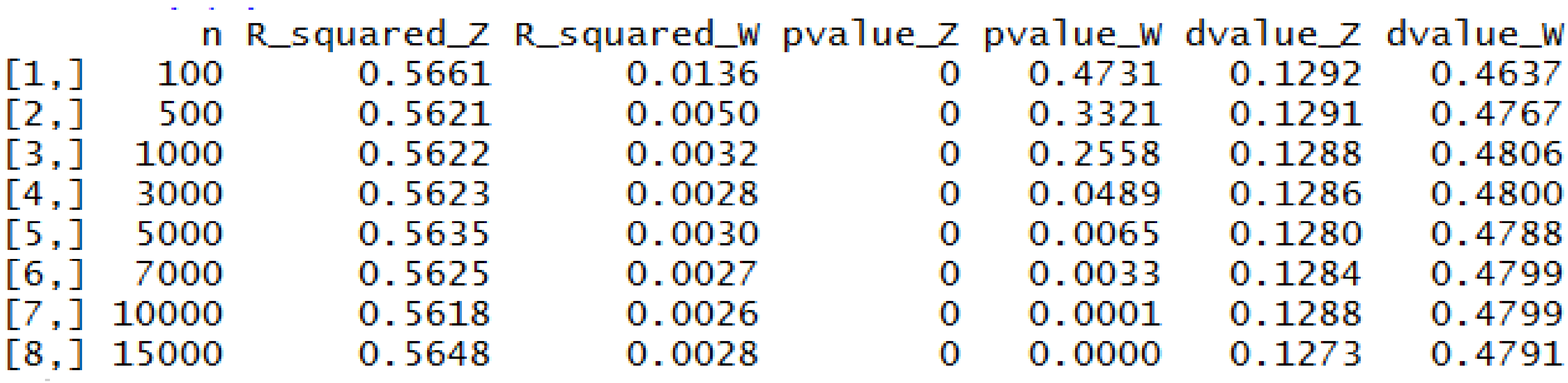
$$\lim_{n\to\infty} \frac{1}{n} \log P_n = \frac{-\beta_1^2}{2\sigma^2/s_x^2}$$

or, $$P_n \approx e^{-n\beta_1^2/2\sigma^2/s_x^2}.$$

Thus, any $\beta_1 \neq 0$ will (eventually) give exponentially small p-values. So, p-value is a measure of sample size in some sense.

Some simulation results -

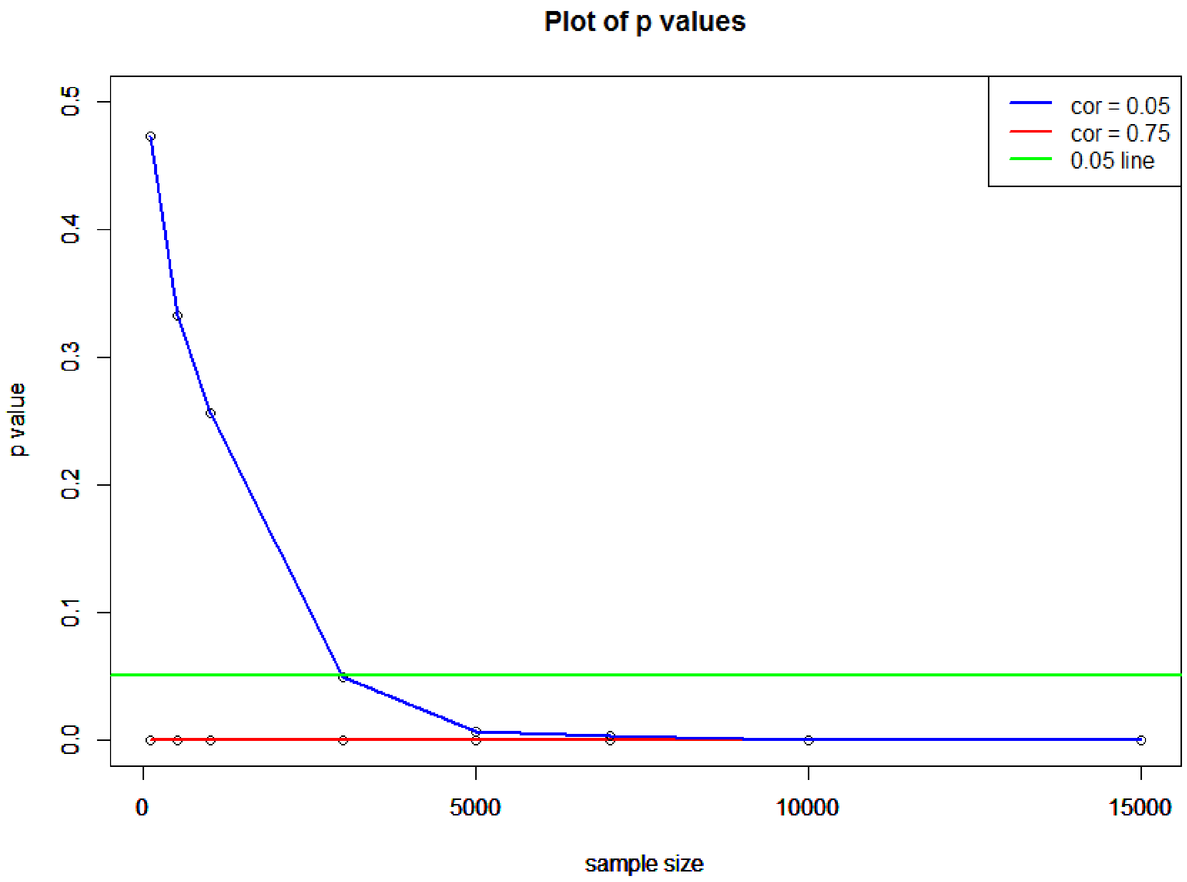
- We generate $X, W \sim N(0,1)$
- Such that $cor(X, W) = 0.05$
- Now we do this for various sample sizes and plot the p value and d value.



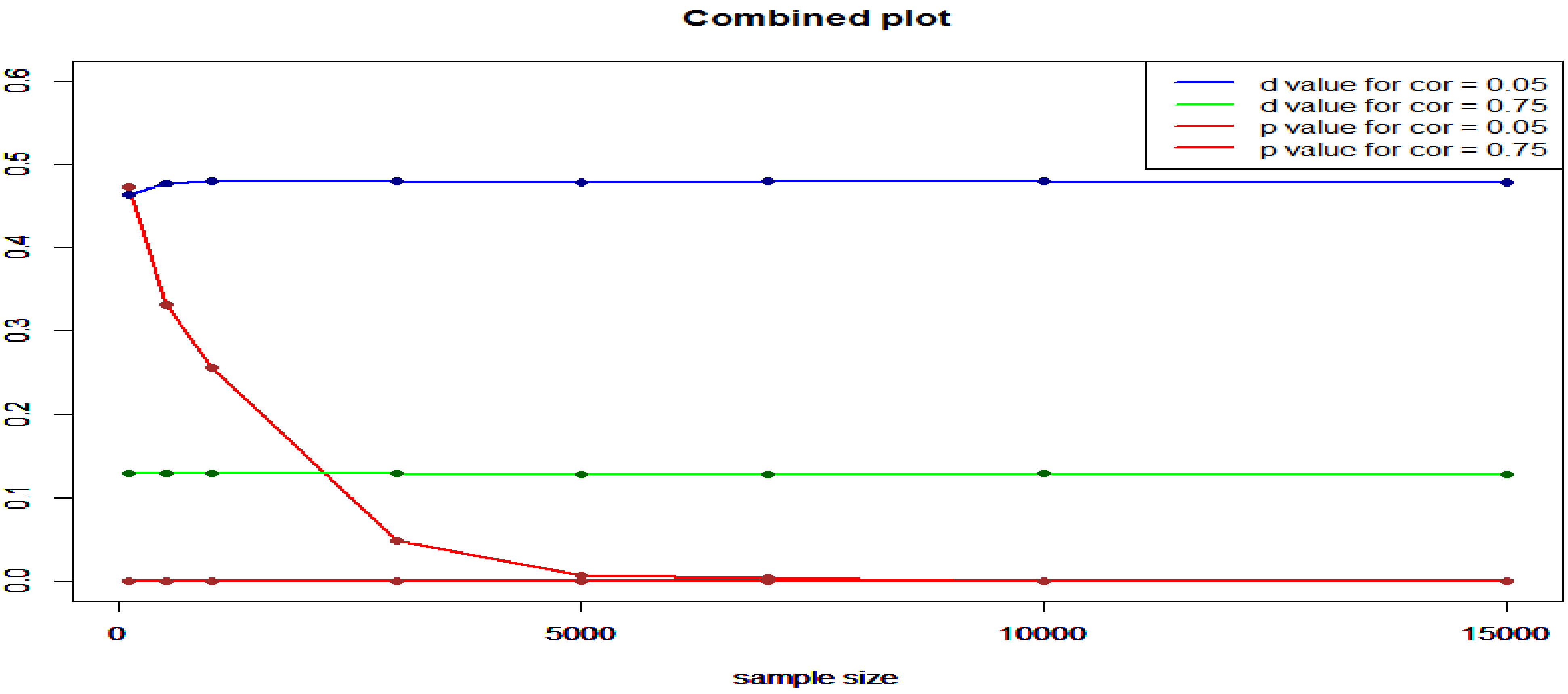
```
          n R_squared_Z R_squared_W pvalue_Z pvalue_W dvalue_Z dvalue_W
[1,]    100      0.5661      0.0136        0   0.4731   0.1292   0.4637
[2,]    500      0.5621      0.0050        0   0.3321   0.1291   0.4767
[3,]   1000      0.5622      0.0032        0   0.2558   0.1288   0.4806
[4,]   3000      0.5623      0.0028        0   0.0489   0.1286   0.4800
[5,]   5000      0.5635      0.0030        0   0.0065   0.1280   0.4788
[6,]   7000      0.5625      0.0027        0   0.0033   0.1284   0.4799
[7,]  10000      0.5618      0.0026        0   0.0001   0.1288   0.4799
[8,]  15000      0.5648      0.0028        0   0.0000   0.1273   0.4791
```

Some simulation results -

- P value decreases with n and after n=4000 the graph lies below 0.05 line.
- That means after n=4000 we will reject the hypothesis for W.
- So for large sample sizes we reject H_0 irrespective of whether there is significant regression effect or not.



Some simulation results –



Summary

- Advantages of D value over p value
  - D value solves the problem of dependence on n (sample size)
  - D value has a clear interpretation as the proportion of people who got worse after treatment.
  - As d value is expressed on probability scale it can be used for other applications like personalized medicine etc.

Future Scope

- Another limitation of p value is its non reproducibility.
- That means p value varies a lot with different samples.
- This problem is not addressed by d value.

Thank You