Computations of statistical power in R

Sharad Silwal
ssilwal@radford.edu
Department of Mathematics and Statistics
Radford University
Radford, VA 24142
USA

Abstract

Hypothesis testing is one of the foundational topics in statistics and a must teach in every introductory statistics course. The entire theory of hypothesis testing rides on the concept of power. Yet, its computation is generally perceived to be daunting and is avoided by most introductory courses in statistics. For instance, it is not covered in AP Statistics curriculum. Understanding of power, on the other hand, is reinforced when you know how to compute power against an alternate hypothesis. This paper presents an exposition of computations and illustrations of the critical value and statistical power for a t-test and its nonparametric analog in R.

1 Introduction

Hypothesis testing is the central tool in inferential statistics where the goal is to confirm or refute a research hypothesis about a population through the observation of a sample. Modern hypothesis testing combines the theory of p-value developed by Ronald Fisher [3] and the theory of hypothesis testing developed by Jerzy Neyman and Egon Pearson [5] in the 30's. It is so widely used now that it forms the bedrock of current scientific research and is included in the syllabus of any first course in statistics.

Hypothesis testing begins with a research hypothesis, also called an alternate hypothesis H_a, constituted by forming a comparison of a population parameter, say, the population mean μ, with some fixed hypothesized value, say, μ_0 using an inequality statement. Next, the probability of making a type-I error, called a significance level, is set, most popularly at $\alpha = 0.05$. This is simply setting a tolerance level that it would be acceptable if the research hypothesis is wrongly confirmed 5% of the time. The alternate hypothesis of $\mu \neq \mu_0, \mu < \mu_0$ and $\mu > \mu_0$ are called the two-tailed, left-tailed, and right-tailed cases respectively. The null hypothesis H_0, on the other hand, is simply $\mu = \mu_0$ which is assumed on the population for the purpose of carrying out the test. Under this assumption, the probability of the observed sample statistic, namely, the sample mean $\bar{x}$, of being the most conservative value to favor the research hypothesis is calculated and is called the p-value. The research hypothesis is confirmed if p-value $< \alpha$ and refuted otherwise.

The decision-making criterion can be put in two different ways. When we make a rule of "reject H_0 if p-value < 0.05", we are essentially saying that the evidence did not favor H_a ($H_a : \mu > \mu_0$ for a right-tailed case) at significance level $\alpha = 0.05$. The evidence is found in the location of the observed $\overline{x}$, let us denote $\overline{x}_{obs}$, in the distribution of $\overline{x}$ under H_0 with $\mu = \mu_0$. Specifically, it is referring to the fact that, under H_0 with $\mu = \mu_0$, the $p-$value quantifying the probability of $\overline{x}$ being at least as large (for a right-tailed case) as $\overline{x}_{obs}$ stayed strictly lower than the significance level of $\alpha = 0.05$. An alternate criterion of making a decision uses the so-called critical value which is the 95^{th} percentile (corresponding to $1 - \alpha = 0.95$ for a right-tailed case) of the sampling distribution of $\overline{x}$ under H_0 with $\mu = \mu_0$. The critical value, let us denote $\overline{x}_c$, thus becomes the cut-off point for the sample mean $\overline{x}$ in order to exhibit any evidence in support of $H_a : \mu > \mu_0$. Thus, an alternate phrasing of the decision-making criterion is "reject H_0 if $\overline{x}_{obs} > \overline{x}_c$".

The whole process of hypothesis testing is founded on the assumption that the null hypothesis H_0 is true which may or may not be the case. So, when a decision is made one way or the other, it falls into one of the four possible scenarios - of which two are correct and the remaining two erroneous, as illustrated in Table 1. When the foundational assumption of H_0 being true is indeed true, then a decision of rejecting H_0 (a positive decision that is in favor of the research hypothesis H_a) is a false positive. This is called the type-I error. On the other hand, when the foundational assumption is false, then a decision of not rejecting H_0 (a negative decision that is against the research hypothesis H_a) is a false negative. In line with its double negative connotation, this error is called the type-II error.

A significance level of $\alpha = 0.05$ means that the probability of not making a type-I error is $1 - 0.05 = 0.95$. This is setting the standard of being able to correctly refute the research hypothesis at least 95% of the time. A type-II error, on the other hand, is the error of incorrectly refuting a research hypothesis. Now, a probability, say, $\beta = 0.2$ of making a type-II error translates to setting the standard of being able to correctly confirm the research hypothesis at most 80% of the time since $1 - \beta = 0.8$ which, in turn, is called the power of the test. It refers to the power of not failing to detect a confirmed instance of the research hypothesis. Naturally, a hypothesis test should aim to have the highest possible power.

Truth / Decision	H_0 is true	H_0 is false
Reject H_0	Type-I error	Correct
Do not reject H_0	Correct	Type-II error

Table 1: Illustration of type-I and -II errors

Unlike the p-value, the power $1 - \beta$ is not as well-understood and certainly not as widely used by early learners of statistics. This is evident in the fact that many first-level statistics courses either do not cover power at all or cover merely its definition without computations. For instance, AP Statistics curriculum doesn't require one to learn how to compute power of a hypothesis test [4]. Yet, hypothesis tests reported in scientific literature do make a point of discussing power of the involved tests.

Sample size determination, an essential component of experimental studies, also requires power calculations. Power is also based on the so-called effect size which, in the case of testing a hypothesized value of the mean, is simply the difference of the means in the null and alternative

hypotheses. Power analysis used in research planning is the determination of the sample size necessary to attain a specified power to detect a hypothesized effect size [2]. Since a bigger sample always guarantees a better power, before proceeding with a costly experimental study, it is prudent to find out what size of a sample one should employ. This is crucial in making the study effective as well as economical. Power and sample size analysis is undertaken by first conducting an inexpensive pilot or simulation study to gauge what sort of an effect size one could anticipate.

This paper illustrates computing power of a one-sample t-test and its non-parametric analog, namely, Wilcoxon signed rank test for a normally distributed alternate. We use the open-source programming language R [6] to compute power analytically. R codes and a plot are provided to make the paper self-contained and reader-friendly. R being the most widely used tool of statistical computing and graphics, readers should be able to easily reproduce all the results provided in this paper.

2 Power of a hypothesis test

Let us look at a dataset of wind speeds of New York, May to September 1973 drawn from the data set New York Air Quality Measurements [1] available in the R datasets package: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html. This dataset has mean 9.557516 mph which, let us suppose, is the maximum wind speed proposed for a certain recreational activity of interest and we would like to know whether or not the wind speed exceeds this desired limit. In order to illustrate this testing with the help of a simple dataset, we will modify the above wind dataset to have mean 0 and standard deviation 1 as follows:

```
> wind <- airquality$Wind - mean(airquality$Wind)
> wind <- wind/ sd(airquality$Wind)
> hist(wind)
> mean(wind)
> sd(wind)
> x <- sample(na.omit(wind),size=27)
> mean(x)
[1]  0.4461874
> sd(x)
[1]  1.064263
```

The wind dataset we created is not exactly normal but does not seem to deviate too much from normality as shown by the histogram created by the `hist` function in R. The R functions `mean` and `sd` compute the mean and standard deviation of the data respectively. The wind dataset, by design, has mean 0 and standard deviation 1. Treating this dataset as our population, we are taking a sample of size 27 from it using the R function `sample`. The research hypothesis we seek to test in this scenario is $H_a : \mu > 0$ at significance level $\alpha = 0.05$. We begin

the hypothesis testing process by making an assumption that the null hypothesis $H_0 : \mu = 0$ holds true.

Now, the power of the testing $H_0 : \mu = 0$ vs. $H_a : \mu > 0$ can be computed for any specific case of the alternate hypothesis $H_a : \mu > 0$. For the sake of convenience of illusration, let us choose the alternate $H_a : \mu = 0.4$, assuming that the true population is normally distributed with mean 0.4 and standard deviation 1. Then our sample $x_1, x_2, \ldots, x_{27}$ are independently and identically distributed values from $N(0.4, 1)$. It is important to note, however, that, in practice, the actual population distribution remains unknown as we proceed to conduct a hypothesis test to confirm or refute $H_a : \mu > 0$.

We shall use two different tests for this testing and compute a power for each. A Wilcoxon signed rank test, a nonparametric test, is a test of choice when the sample $x_1, x_2, \ldots, x_{27}$ does not display any normality. If, on the other hand, the sample displays normality, one would choose a t-test for this purpose.

3 Power of a Wilcoxon signed rank test

A Wilcoxon signed ranked test uses ranks instead of actual values and a median instead of a mean. So, we seek to test Med > 0 on the sample $x_1, x_2, \ldots, x_{27}$. Since the hypothesized median value or the null hypothesis is Med $= 0$, the ranks of 1 through 27 are determined for the values $|x_i - 0| = |x_i|$.

The Wilcoxon signed rank test statistic is given by

$$W = \sum_{i=1}^{27} W_i := \sum_{i=1}^{27} Z_i R_i,$$

where R_i is the rank of $|x_i|$ and $Z_i = 0$ if x_i is negative and $Z_i = 1$ if x_i is positive. This means each W_i is either 0 or some k where $k = 1, \ldots, 27$ as prescribed by the distribution of x_i's, that is, according to whether x_i is negative or positive. For the purpose of computing the expected value and the variance of W, denoted $E(W)$ and $Var(W)$ respectively, without loss of generality, we can reorder either x_i's or W_i's so that each W_i is either 0 or exactly i according to the same distribution of x_i's.

For $n = 27$, W could be anything between 0 and $\frac{27(27+1)}{2} = 378$. From a table of percentiles of W, we see that the actual critical value for a left-tailed case for $\alpha = 0.05$ is 120 and, for a right-tailed case, H_0 is rejected if $W \geq (378 - 120) = 258$. We can also find an approximate normal distribution of W.

By the Central Limit Theorem, W, under the null Med $= 0$, approximately follows a normal distribution with mean $\frac{27(27+1)}{4} = 189$ and standard deviation $\sqrt{\frac{27(27+1)(2\cdot 27+1)}{24}} = \sqrt{\frac{3465}{2}} = \sqrt{1732.5} = 41.62$. We can prove this as follows.

Note that under $H_0 :$ Med $= 0$, the symmetry of the discrete probability distribution of W_i yields

$$p(W_i = 0) = 0.5, p(W_i = i) = 0.5, p(W_i^2 = 0) = 0.5 \quad \text{and} \quad p(W_i^2 = i^2) = 0.5.$$

Then we can compute

$$E(W) = \sum_{i=1}^{27} E(W_i) = \sum_{i=1}^{27} [0 \cdot 0.5 + i \cdot 0.5] = 0.5 \sum_{i=1}^{27} i$$
$$= (0.5) \cdot \frac{27(27+1)}{2} = \frac{378}{2} = 189.$$

$$Var(W) = \sum_{i=1}^{27} Var(W_i) = \sum_{i=1}^{27} [E(W_i^2) - E(W_i)^2]$$
$$= \sum_{i=1}^{27} \left[0^2 \cdot 0.5 + i^2 \cdot 0.5\right] - (i/2)^2 = \frac{1}{4} \sum_{i=1}^{27} i^2$$
$$= \frac{1}{4} \cdot \frac{27(27+1)(2 \cdot 27+1)}{6} = \frac{3465}{2}.$$

Next, for the alternate H_a : Med $= 0.4$, we can compute the probability of a random value from the true population being negative in R as follows:

```
> pnorm(0,mean=0.4,sd=1)
[1]   0.3445783
```

Using the normal cumulative density function `pnorm`, this computation found 0 to be the 34.46^{th} percentile in the normal distribution with mean 0.4 and standard deviation 1. So, the discrete probability distribution of W_i, under the alternate Med $= 0.4$, yields

$$p(W_i = 0) = p(W_i^2 = 0) \approx 0.3446 \quad \text{and} \quad p(W_i = i) = p(W_i^2 = i^2) \approx 0.6554.$$

Then we can compute

$$E(W) = \sum_{i=1}^{27} E(W_i) = \sum_{i=1}^{27} [0 \cdot 0.34 + i \cdot 0.66] = 0.66 \sum_{i=1}^{27} i$$
$$= (0.66) \cdot \frac{27(27+1)}{2} \approx 247.7412.$$

$$Var(W) = \sum_{i=1}^{27} Var(W_i) = \sum_{i=1}^{27} [E(W_i^2) - E(W_i)^2]$$
$$= \sum_{i=1}^{27} \left[0^2 \cdot 0.34 + i^2 \cdot 0.66\right] - (0.66 \cdot i)^2 = (0.66 - 0.66^2) \sum_{i=1}^{27} i^2$$
$$= (0.2259) \cdot \frac{27(27+1)(2 \cdot 27+1)}{6} \approx 1565.487.$$

Noting $\sqrt{1565.487} \approx 39.5662$, we now can compute the power of Wilcoxon signed rank test for the alternate Med $= 0.4$ as follows:

```
> qnorm (0.05, mean=189, sd=41.62, lower.tail=FALSE)
[1]  257.4588
> pnorm (257.4588, mean=247.7412, sd=39.5662, lower.tail=FALSE)
[1]  0.4029946
```

The normal quantile function `qnorm` is the inverse of the normal cumulative density function `pnorm` in R. This computation showed that the 95^{th} percentile in the normal distribution with mean 189 and standard deviation 41.62 is 257.4588. Hence, in testing H_0 : Med $= 0$ vs. H_a : Med > 0, the power of a Wilcoxon signed rank test for the alternate Med $= 0.4$ is 0.40.

4 Power of a t-test

If the sample does not violate normality test, one would opt for a t-test. Note that, for a sample size of $n = 27$, our sample standard error is $1.06/\sqrt{27} \approx 0.2040$ and using the `ggdist` package in R for scaled and shifted t-distributions, we can compute the power as follows:

```
> library(ggdist)
> qstudent_t(0.05, df=26, mu=0, sigma=0.2040, lower.tail=FALSE)
[1]  0.3479461
> pstudent_t(0.3479461, df=26, mu=0.4, sigma=0.2040,
lower.tail=FALSE)
[1]  0.5996988
```

Hence, in testing $H_0 : \mu = 0$ vs. $H_a : \mu > 0$, the power of a t-test for the alternate $\mu = 0.4$ is 0.60. Note that a t-test produced a better power than a nonparametric test as we would expect due to the normality of the data.

It must be noted that the sample standard deviation varies from sample to sample, staying close enough to the population standard deviation for a large enough sample size. So, let us also investigate scenarios where the sample standard deviation is equal to or less than the population standard deviation. If we use $s = \sigma = 1$, the sample standard error is $1/\sqrt{27} \approx 0.1925$ and the computation goes as follows:

```
> library(ggdist)
> qstudent_t(0.05, df=26, mu=0, sigma=0.1925, lower.tail=FALSE)
[1]  0.3283314
> pstudent_t(0.3283314, df=26, mu=0.4, sigma=0.1925,
lower.tail=FALSE)
[1]  0.6436581
```

Hence, in the case of $s = \sigma = 1$, the power of a t-test for the alternate $\mu = 0.4$ in testing $H_a : \mu > 0$ is 0.64.

If we use $s = 0.99$, the sample standard error becomes $0.99/\sqrt{27} \approx 0.1905$ and the computation goes as follows:

```
> library(ggdist)
> qstudent_t(0.05, df=26, mu=0, sigma=0.1905, lower.tail=FALSE)
[1]  0.3249202
> pstudent_t(0.3249202, df=26, mu=0.4, sigma=0.1905,
lower.tail=FALSE)
[1]  0.6516467
```

Thus, in the case of $s = 0.99 < 1 = \sigma$, the power of a t-test for the alternate $\mu = 0.4$ in testing $H_a : \mu > 0$ is 0.65.

5 Power of a z-test

From the vantage point of knowing that our population is indeed normally distributed, let us find the power of a z-test as well. We should see a still better power for a z-test. Noting the population standard error is $1/\sqrt{27} \approx 0.1925$, the computation in R goes as follows:

```
> qnorm(0.05, mean=0, sd=0.1925, lower.tail=FALSE)
[1]  0.3166343
> pnorm (0.3166343, mean=0.4, sd=0.1925, lower.tail=FALSE)
[1]  0.6675175
```

As expected, a z-test has the best power of 0.67 for this testing. Its illustration in R is provided in Figure 1.

R codes used to generate Figure 1 are provided below:

```
> library(ggplot)
> library(reshape2)
> crit <- qnorm (0.05, mean=0, sd=0.1925, lower.tail=FALSE)
> power <- pnorm (0.3166343, mean=0.4, sd=0.1925, lower.tail=FALSE)
> x <- seq(-1, 1, length = 100)
> H0 <- dnorm(x,0,0.1925)
> Ha <- dnorm(x,0.4,0.1925)
> normdist <- data.frame(cbind(x,H0,Ha))
> colnames(normdist) <- c("x","H0","Ha")
> normdist <- melt(normdist,id=c("x"))
```

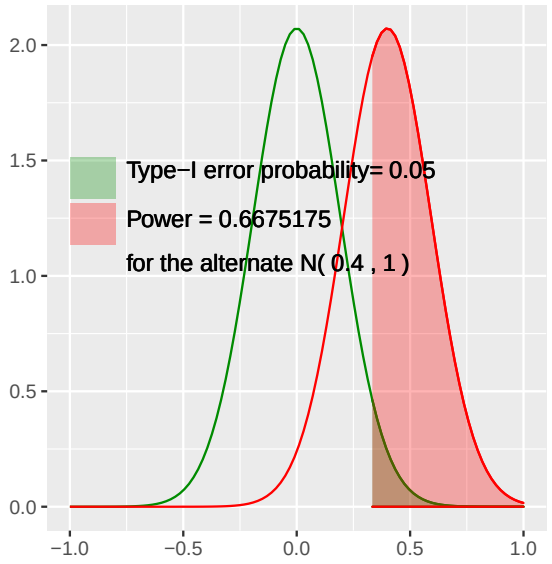

Figure 1: Illustration of statistical power.

```
> g <- ggplot(normdist,aes(x, value, color=variable))+
+   scale_color_manual(values=c("H0" = "green4","Ha" = "red"))+
+   geom_line(size=.5) + theme(axis.title=element_blank())+
+ theme(legend.position = "none") + geom_ribbon(data
=subset(normdist,x>crit), aes(x=x, ymax=value,
group=variable,fill=variable),ymin=0,alpha=0.3)+
+ scale_fill_manual(name =` `, values = c("H0" = "green4", "Ha" =
"red"))
```

6 Power functions

In the foregoing sections we computed statistical powers of three different hypothesis tests, all testing whether or not the average is positive, under an alternate of the true population being normally distributed with mean $\mu = 0.4$ and standard deviation $\sigma = 1$. In this section we illustrate some computations in power analysis.

Let us compare the differences in proximity of the three hypothesis tests that we used with our true population distribution. Comparing the assumptions of each hypothesis test with the reality of the true population, we find that the z-test coincides with the true population distribution, the t-test is some distance away, and the Wilcoxon signed rank test comes the furthest away. The Wilcoxon signed rank test had the power of 0.40 while the power of the z-test was 0.67. When the sample standard deviation s was 1.06, 1, and 0.99 (different comparisons with the population standard deviation of $\sigma = 1$), the power of the t-test was 0.60, 0.64, and 0.65 respectively. The sample standard deviation s is an estimator of, and thus should stay fairly close to, the population standard deviation α. However, when s is significantly smaller than σ, particularly in cases of small sample sizes such as n=27 in our example, it is possible for

a t-test to have a better power than a z-test. But then again the true population distribution is more likely to be a t-distribution than a normal distribution in such cases. Hence, we can safely conclude from our observations that, with other things being equal, the closer the assumptions of the hypothesis test with the true scenario, the better the power.

To understand how power behaves as functions of some of the parameters of power analysis, let us restrict ourselves to the case of a z-test. The first parameter of interest is the effect size, or the mean of the alternate true population. If we change it to $\mu = 0.5$, we can compute the following:

```
> pnorm (0.3166343, mean=0.5, sd=0.1925, lower.tail=FALSE)
[1]  0.8295907
```

This increased power from 0.67 to 0.83. Hence, power as a function of the effect size is an increasing function.

If we change the sample size from 27 to 29, the population standard error becomes $1/\sqrt{29} \approx 0.1857$, and we compute

```
> qnorm(0.05, mean=0, sd=0.1857, lower.tail=FALSE)
[1]  0.3054493
> pnorm (0.3054493, mean=0.4, sd=0.1857, lower.tail=FALSE)
[1]  0.6883478
```

Thus, power increased from 0.67 to 0.69 as the sample size increased from 27 to 29. This illustrates that power will go up when a larger sample size is adopted.

Let us now change the type-I error probability or the significance level α from 0.05 to 0.01 and compute

```
> qnorm(0.01, mean=0, sd=0.1925, lower.tail=FALSE)
[1]  0.447822
> pnorm (0.447822, mean=0.4, sd=0.1925, lower.tail=FALSE)
[1]  0.4019024
```

Lowering the significance level from 0.05 to 0.01 brought the power down from 0.67 to 0.40. Hence, we observe that power as a function of the type-I error probability α is an increasing function. Since power is the complement of the type-II error probability β, this shows a trade-off between the type-I and type-II error probabilities. This means it is not possible to control both types of error simultaneously although that is exactly what is in our interest. As such, the popular convention is to keep the type-I and type-II error probabilities at 0.05 and 0.20 respectively. In other words, a significance level of $\alpha = 0.05$ and a power of $1 - \beta = 0.8$ have become the gold standard of hypothesis testing in practice.

Every student who is introduced to hypothesis testing must learn the computations illustrated above in this section. Computing powers corresponding to several values of a test parameter allows us to graph the power function and visualize the dependence of power on the

parameter. These computation exercises impart tangible insight into how power is related to effect size, sample size, and significance level. This approach to teaching hypothesis testings reinforces students' knowledge of p-values and the two types of error. Since the p-value is directly related to the significance level as discussed in the introduction section, students will be able to see p-values in relation to power as opposed to the conventional practice of only looking at the p-values.

7 Conclusion

Powers of a Wilcoxon signed rank test, a t-test and a z-test along with their R codes and an illustration were presented in this paper. A specific normal distribution was used as an alternate with an effect size of 0.4. In particular, it was shown how power is compromised when a nonparametric test is used when the population is known to be normally distributed.

Statistical power is a very practical tool and should be taught in every introductory-level statistics class just when hypothesis testing is being introduced. A good working knowledge of power will facilitate a solid understanding of the theory of hypothesis testing. This, however, should mean learning not just the definition of power as is the current practice but also its computations in specific scenarios. Exposure to power is an effective way of reducing students' anxiety about the elusive p-value. Statistical power harmoniously ties in with the concepts of p-value and the two types of errors and, learned together, they reinforce each other's comprehension in students.

References

[1] Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983). Graphical Methods for Data Analysis. Belmont, CA: Wadsworth.

[2] Cohen, Jacob. (2013). Statistical power analysis for the behavioral sciences. Routledge.

[3] Fisher, R. A. (1925). Statistical Methods for Research Workers. (Oliver and Boyd, Edinburgh, UK).

[4] Retrieved from https://apstudent.collegeboard.org/apcourse/ap-statistics/course-details.

[5] Neyman, J. and Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231(694-706), 289-337.

[6] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.