Name_____ **Date**_____ **Class**_____

WISE Power Tutorial – All Exercises

Power: The B.E.A.N. Mnemonic

Four interrelated features of power can be summarized using **BEAN**

B Beta Error (Power = 1 – Beta Error): Beta error (or Type II error) is the probability that a test of statistical significance will fail to reject the null hypothesis when it is false (e.g., when there really is an effect of training).

- As beta error increases, power decreases.

E Effect Size: The effect size is the magnitude of the difference between the actual population mean and the null hypothesized mean ($\mu_1 - \mu_0$) relative to standard deviation of scores (σ). When the effect size $d = (\mu_1 - \mu_0) / \sigma$ in the population is larger, the null and population sampling distributions overlap less and power is greater.

- As effect size increases, power increases (assuming no change in alpha or sample size).

A Alpha error: Alpha error (or Type I error) is the probability that a statistical test will produce a statistically significant finding when the null hypothesis is true (e.g., there is no effect of training). For example, if alpha error = .05 and the null hypothesis is true, then out of 100 statistical tests, false significance would be found on average 5 times. The risk of false significance would be 5%. In practice, alpha is typically set by the researcher at .01 or .05

- As alpha error increases, power increases (assuming no change in effect size or sample size).

N Sample Size: As the sample size increases, the variability of sample means decreases. The population and null sampling distributions become narrower, overlapping to a lesser extent and making it easier to detect a difference between these distributions. This results in greater power.

- As sample size increases, power increases (assuming no change in alpha or effect size).

Select true or false for each scenario:

(Assuming no other changes)	True	False
1. As effect size increases, power decreases.	☐	☐
2. As sample size increases, power increases.	☐	☐
3. As alpha error increases, power decreases.	☐	☐
4. Beta error is unrelated to power.	☐	☐

1. Power and Effect Size

As the effect size increases, the power of a statistical test increases. The effect size, *d*, is defined as the number of standard deviations between the null mean and the alternate mean. Symbolically,

$$d = \frac{\mu_1 - \mu_0}{\sigma}$$

where *d* is the effect size, μ_0 is the population mean for the null distribution, μ_1 is the population mean for the alternative distribution, and σ is the standard deviation for both the null and alternative distributions. From the equation it can be noted that two factors impact the effect size: 1) the difference between the null and alternative distribution means, and 2) the standard deviation. For any given population standard deviation, the greater the difference between the means of the null and alternative distributions, the greater the power. Further, for any given difference in means, power is greater if the standard deviation is smaller. In the following exercise, we will use the power applet to explore how the effect size influences power. Your task is to find a good way to explain how this works to a friend.

Exercise 1a: Power and Mean Differences (Large Effect)

How probable is it that a sample of graduates from the ACE training program will provide convincing statistical evidence that ACE graduates perform better than non-graduates on the standardized Verbal Ability and Skills Test (VAST)? How likely is it that a rival competitor, the DEUCE training program, will provide convincing evidence? Power analysis will allow us to answer these questions.

We will use the WISE Power Applet to examine and compare the statistical power of our tests to detect the claims of the ACE and DEUCE training programs. We begin with a test of ACE graduates.

We assume that for the population of non-graduates of a training course, the mean on VAST is 500 with a standard deviation of 100. For the population of ACE graduates the mean is 580 and the standard deviation is 100. Symbolically, $\mu_0 = 500$, $\mu_1 = 580$, and $\sigma = 100$. Both distributions are assumed to be normal.

How large is the effect size? The formula for *d* shown below indicates that the effect size for the ACE program is .80. This tells us that the mean for the alternative population is .80 standard deviations greater than the mean for the null population.

$$d = \frac{\mu_1 - \mu_0}{\sigma} = \frac{580 - 500}{100} = .80$$

The z-score of a sample mean computed on the null sampling distribution allows us to determine the probability of observing a sample mean this large or larger if the null hypothesis is true.

To prepare the simulation, enter the following information into the applet below:

- μ_0 = 500 (**null mean**);
- μ_1 = 580 **(alternative mean**);
- σ = 100 (**standard deviation**);
- α = .05 (**alpha error rate, one tailed**);
- **n** = 25 (**sample size**).

To simulate drawing one sample of 25 cases, press **Sample**. The mean and *z*-score are shown in the applet (bottom right box). Record these values in the first pair of boxes below (you may round the mean to a whole number).

Trial	1	2	3	4	5	6	7	8	9	10
Mean				579	574	594	600	541	585	578
Z-Score				3.96	3.72	4.69	4.99	2.04	4.23	3.92

Is this sample mean large enough to allow you to reject the null hypothesis? How likely is it that you would observe a sample this large or larger if the null hypothesis was true so that you really were sampling from the blue distribution? (Answer: The p-value is the probability of observing a mean as large or larger than your sample mean if the null hypothesis is true.)

Now draw two more samples and record the mean and z for each in the boxes. These values will be saved and used later and can be printed for a homework exercise. Some of the boxes have already been filled out for you.

The power of this statistical test is the probability that the mean of a random sample of size n will be large enough to allow us to correctly reject the null hypothesis. Because we are actually sampling from the **Alternative Population** (red distribution), the probability that we will observe a sample mean large enough to reject H_0 corresponds to the proportion of the red sampling distribution that is to the right of the dashed line. For this example, we can use the value provided by the applet, **.991**.

Thus, if we draw a sample of 25 cases from ACE graduates, the probability is 99.1% that our sample mean will be large enough that we can reject the null hypothesis that the sample came from a population with a mean of only 500. The probability that we will fail to reject H_0 is only 1.000 - .991 = .009, less than one chance in 100.

1a. How many times could you reject the null hypothesis in your ten samples?
(With one-tailed alpha $\alpha = .05$, $z = 1.645$, so reject H_0 if your z-score is greater than 1.645)

Exercise 1b: Power and Mean Differences (Small Effect)

Now we will assess the power of a test for a rival training program, the DEUCE program. The mean score for the population of graduates of this program is 520. Again we assume the population distribution is normal with a standard deviation of 100. Using the formula for *d*, we find that the population effect size for the DEUCE program is only .20.

$$d = \frac{\mu_1 - \mu_0}{\sigma} = \frac{520 - 500}{100} = .20$$

Recall the effect size for the ACE program was much larger:

$$d = \frac{\mu_1 - \mu_0}{\sigma} = \frac{580 - 500}{100} = .80$$

1b. Before drawing samples, consider how the statistical power will differ for a test of DEUCE graduates compared to the power we found for a test of ACE graduates. That is, do you expect you will be more likely or less likely to reject the null hypotheses for a sample of 25 graduates drawn from the DEUCE program compared to a similar test for the ACE program?

I predict that statistical power for the test of the DEUCE program compared to the test of the ACE program will be:

Less The Same Greater

With the applet you will be able to change the effect size and watch what happens to statistical power. Your goal for this exercise is to be able to explain to a friend why statistical power is greater when the effect size is greater.

To simulate drawing a sample of 25 from graduates from the DEUCE program, enter the following information into the WISE Power Applet:

- μ_0 = 500 (**null mean**);
- μ_1 = 520 **(alternative mean**);
- σ = 100 (**standard deviation**);
- α = .05 (**alpha error rate, one tailed**);
- **n** = 25 (**sample size**).

Do three simulations of drawing a sample of 25 cases, and record the results below.

Trial	1	2	3	4	5	6	7	8	9	10
Mean				509	511	513	502	492	513	533
Z-Score				0.44	0.54	2.06	0.11	-0.41	0.65	1.63

1c. What is the power for this test as shown in the applet?

1d. How many of your ten simulated samples allowed you to reject the null hypothesis?

(Use one-tailed alpha α = .05, z = 1.645, so reject H_0 if your z-score is greater than 1.645)

1e. For the ACE program, the effect size was **.8** and the power of the statistical test was **.991**; what can you conclude about the relationship between effect size and power?

- **A.** The test for the ACE program, which had a larger effect size, had more power.
- **B.** The test for the DEUCE program, which had a smaller effect size, had more power.
- **C.** Effect size is unrelated to power.

Exercise 1c: Power and Variability (Standard Deviation)

In Exercises 1a and 1b, we examined how differences between the means of the null and alternative populations affect power. In this exercise, we will investigate another variable that impacts the effect size and power; the variability of the population. If the standard deviation for graduates of the TREY program was only 50 instead of 100, do you think power would be greater or less than for the DEUCE program (assume the population means are 520 for graduates of both programs)? Think about what will happen before you try the simulation. Referencing the effect size calculation may help you formulate your opinion:

$$d = \frac{\mu_1 - \mu_0}{\sigma}$$

1f.I think that with a smaller standard deviation in the population, the statistical power will be

☐ Less ☐ Unchanged ☐ Greater ☐ I don't know

To simulate drawing a sample from graduates of the TREY program that has the same population mean as the DEUCE program (520), but a smaller standard deviation (50 instead of 100), enter the following values into the WISE Power Applet:

- μ_0 = 500 (**null mean**);
- μ_1 = 520 **(alternative mean)**;
- σ = 50 (**standard deviation**);
- α = .05 (**alpha error rate, one tailed**);
- **n** = 25 (**sample size**).

Do three simulations of drawing a sample of 25 cases and record the results below.

Trial	1	2	3	4	5	6	7	8	9	10
Mean				512.1	516.4	515.6	515.4	525.2	535.3	528.6
Z-Score				1.21	1.64	1.56	1.36	2.52	3.53	2.86

1g. How many of your ten simulated samples allowed you to reject the null hypothesis? (Use one-tailed alpha $\alpha = .05$, $z = 1.645$, so reject H_0 if your z-score is greater than 1.645) ☐

1h. What is the power for this test (from the applet)? ☐

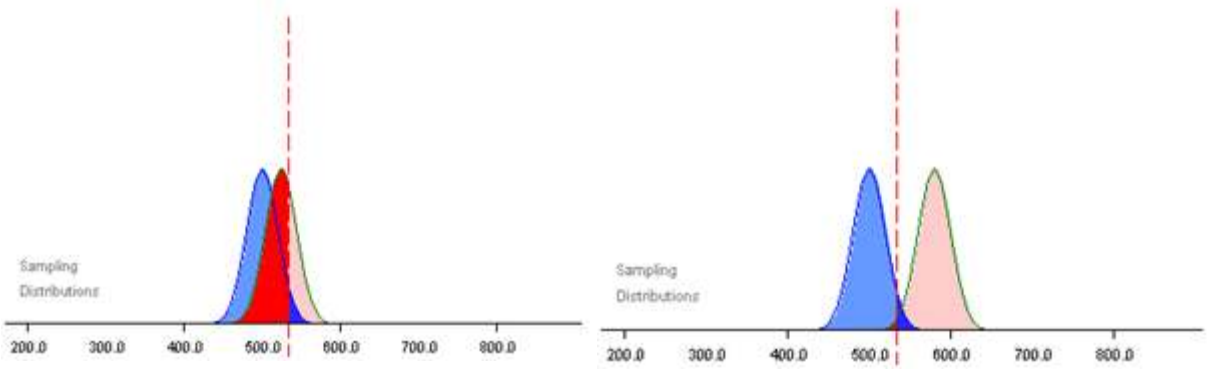
1i. In Exercise 1b the DEUCE program had a mean of 520 just like the TREY program, but with samples of N = 25 for both programs, the test for the DEUCE program had a power of .260 rather than .639. The standard deviation for DEUCE was 100 rather than 50. Why is statistical power greater for the TREY program?

- ☐ Because smaller population variance always produces greater power.
- ☐ Because the program with the larger effect size always produces greater power.
- ☐ Neither of these reasons is sufficient.

Exercise 1d: Summary of Power and Effect Size

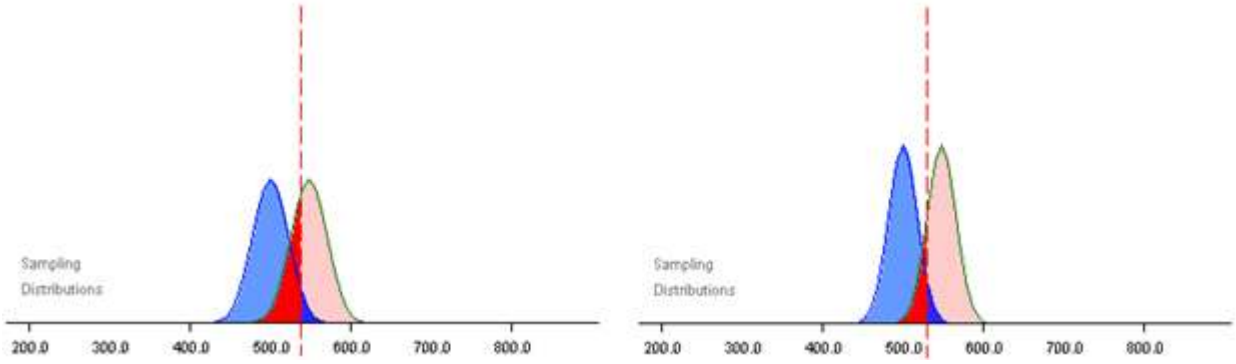
The formula for effect size is $d = \frac{\mu_1 - \mu_0}{\sigma}$. We can see from the formula that two variables impact the effect size: the difference between the means of the alternative and null populations and the standard deviation of the two populations. Below are two images that may be useful for examining the impact of standard deviation and mean differences on power.

Same Standard Deviation, Different Means



When the difference between the means of the null and the alternative distributions is increased, the sampling distributions do not change shape, but they are further apart on the x-axis. Note that when the difference is increased, the sampling distributions overlap to a lesser extent, and it is less likely that a random sample from the alternative distribution will be mistaken for a sample taken from the null distribution. More of the sampling distribution for the alternate population (the pink distribution) is greater than the critical value (red dashed line), so it is more likely that a random sample mean will lead to rejection of the null hypothesis.

Same Mean Difference, Different Standard Deviation



When the standard deviation is smaller, the sampling distributions are narrower, but the means remain the same distance apart on the x-axis. The sampling distributions overlap to a lesser extent because they are narrower. Compare the proportion of the alternate (red) distribution that is above or below the critical value in the two scenarios.

The following questions are designed to test your understanding of the factors that affect power.

Below are key statistics for each of two new training programs, SLAM and DUNK.

Statistics for SLAM	Statistics for DUNK
μ_0 = 500 **(null mean)**; μ_1 = 540 **(alternative mean)**; σ = 50 **(standard deviation)**; α = .05 **(alpha error, one tailed)** n = 50 **(sample size)**.	μ_0 = 500 **(null mean)**; μ_1 = 520 **(alternative mean)**; σ = 20 **(standard deviation)**; α = .05 **(alpha error, one tailed)** n = 50 **(sample size)**.

Do you expect that a test of statistical significance would have greater power for the SLAM program or the DUNK program? Why? Respond to the following true/false statements. See if you can answer all statements correctly before you check your answers.

T	F	
☐	☐	**1j.** The test of the SLAM program will have greater power because SLAM has a larger mean than DUNK.
☐	☐	**1k.** The test of the SLAM program will have greater power because SLAM has a larger standard deviation than DUNK.
☐	☐	**1l.** The test of the DUNK program will have greater power, because a program with smaller standard deviation has greater power.

☐ ☐ **1m.** The test of the SLAM program will have greater power because SLAM has a greater effect size.

☐ ☐ **1n.** The test of the DUNK program will have greater power than the test for the SLAM program if sample size and alpha are the same, because DUNK has a greater effect size.

☐ ☐ **1o.** Power is the same for the tests of the two programs because the samples have the same size.

☐ ☐ **1p.** Holding all other factors constant, a larger difference between the null and alternate population means will always yield greater power.

☐ ☐ **1q.**Holding all other factors constant, power is greater when the variance of the null and alternate populations is greater.

Exercise 2: Power and Sample Size

Why does statistical power increase as sample size increases (assuming effect size and alpha are unchanged)? The goal of this exercise is to help you develop an explanation that you could give to a classmate.

In Exercise 1d we drew samples of 25 graduates from the DEUCE program, but in Exercise 2 we will draw samples of n = 100 and n = 4. Watch what happens and think about how you can explain how statistical power is influenced by sample size.

To simulate drawing a larger sample, enter the following values into the WISE Power Applet:

- μ_0 = 500 (**null mean**);
- μ_1 = 520 **(alternative mean**);
- σ = 100 (**standard deviation**);
- α = .05 (**alpha error rate, one tailed**);
- **n** = 100 (**sample size**).

For this exercise, do ten simulations of drawing a sample of 100 cases and record the results below. You don't need to record means; just select the button for "Reject H_0" or "Fail to Reject" for each of the ten simulations.

First, do ten simulations of drawing a sample of 100 cases and record the results below.

	1	2	3	4	5	6	7	8	9	10
Reject H_0	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Fail to Reject	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

2a. Power (as shown in the applet) = ☐

Next, do ten simulations of drawing a sample of 4 cases and record the results below.

(Set sample size n = 4, press Enter)

	1	2	3	4	5	6	7	8	9	10
Reject H_0	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Fail to Reject	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

2b. Power (as shown in the applet) = ☐

2c. How many times out of 10 did you Reject H_0 for each of the two scenarios?

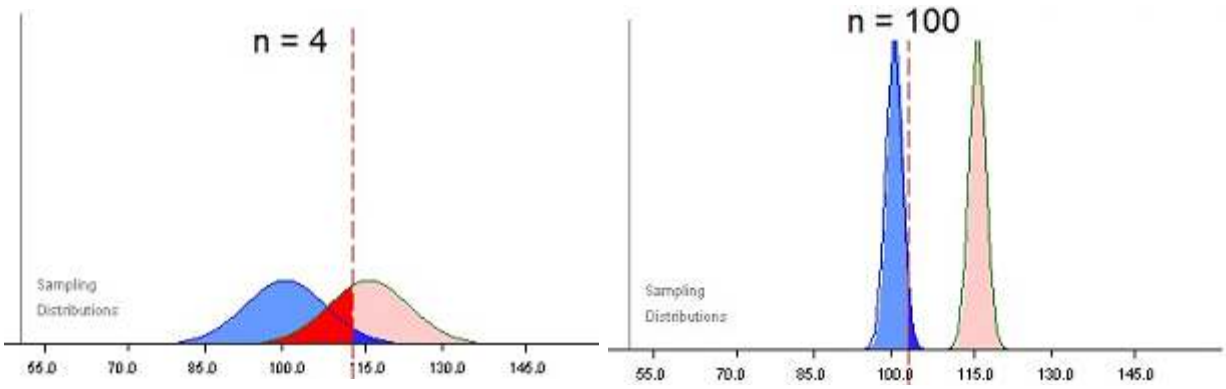
n = 4: ☐
n = 100: ☐

If nothing else is changed, power is greater with a larger sample size because:

(select True or False for each before you check your answers):

T	**F**	
☐	☐	2d. The effect size is larger.
☐	☐	2e. The alpha error is smaller.
☐	☐	2f. The alpha error is larger.
☐	☐	2g. The population variance is smaller.
☐	☐	2h. The variance of the alternate sampling distribution is smaller.
☐	☐	2i. More of the alternate sampling distribution exceeds the critical value.

We have seen that the power of a statistical test increases as the sample increases. Let's review the reasons behind this increase in power. You may have noticed that the sampling distribution of the mean was much wider when the sample size was reduced from 100 to 4. This occurred because sample means are much more variable when the sample size is small. Regardless of sample size, the average sample mean, taken across infinite samples, would be equal to the population mean. However, the precision of sample means is less when the sample size is small. This can be observed as the greater variability in sampling distributions for the sample size of 4. When the sampling distributions are more variable, it is more likely that a sample taken from the alternative distribution will be mistaken for a sample taken from the null distribution. This mistake leads to failure to reject the null hypothesis, which corresponds to lower power. The figures below show sampling distributions for samples of 100 compared to samples of 4. Although the difference between the means is the same, there is much less overlap of the sampling distributions for the larger samples.

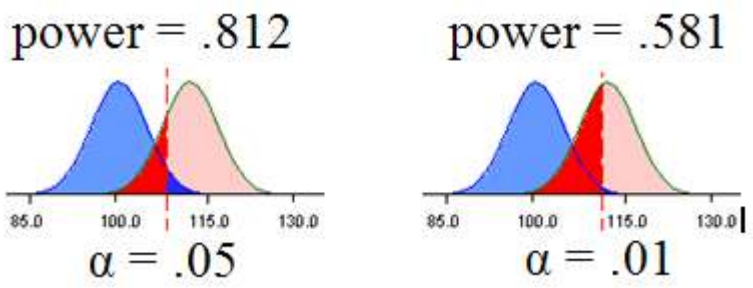


2j. Explain to a classmate why statistical power increases as the sample size increases.

Exercise 3: Power and Alpha

Now, we will consider the impact of using a different alpha value, α.

As the researcher, we decide on the value of alpha, typically at .05 or .01. Alpha is the error rate we are willing to accept for the error of rejecting the null hypothesis if it were true. We require stronger evidence to reject the null hypothesis if we set alpha at .01 than if we use alpha of .05.



The figures show how power and alpha error are related. When we reduce alpha error (represented by the dark blue area of the null distribution to the right of the red dashed line), we also reduce power (the pink area in the alternate distribution to the right of the dashed line).

For this example, use one-tailed alpha $\alpha = .01$ ($z = 2.326$). In this case, we will reject the null hypothesis only if a sample mean is so large that it would occur less than 1% of the time given the null hypothesis is true. You do not need to draw additional samples for this problem; you can use the data recorded for samples drawn in **Exercise 1** ($\mu_0 = 500$, $\sigma = 100$, **n** = 25, $\alpha = .05$, $z = 1.645$).

Data from Exercise 1

ACE Program ($\mu_1 = 580$)

Trial	1	2	3	4	5	6	7	8	9	10
Mean				579	574	594	600	541	585	578
Z-Score				3.96	3.72	4.69	4.99	2.04	4.23	3.92

DEUCE Program ($\mu_1 = 520$)

Trial	1	2	3	4	5	6	7	8	9	10
Mean				509	511	513	502	492	513	533
Z-Score				0.44	0.54	0.65	0.11	-0.41	0.65	1.63

3a. Using alpha of .01 instead of .05, how many times could you reject the null hypothesis for your results in **Exercise 1**? (How many times is Z > 2.326?)

	α = .05 (from #1)	α = .01
Reject for ACE Program (μ_1 = 580)		
Reject for DEUCE Program (μ_1 = 520)		

3b. What is the power for each of these tests? You can use the applet below to calculate power for the tests using alpha α = .01. (Set n = 25 and μ_0 = 500 for all tests ; use μ_1 = 580 for ACE and μ_1 = 520 for DEUCE). Remember to press 'Enter' after each change to the applet.

	α = .05 (from #1)	α = .01
Power for ACE Program (μ_1 = 580)	.991	
Power for DEUCE Program (μ_1 = 520)		

Cumulative Test: What affects Statistical Power?

Select True or False for each of the following questions.

If nothing else is changed, power is greater when…

	T	F	
C1.	☐	☐	The difference between the null and alternative population means is greater.
C2.	☐	☐	The standard deviation of the populations is greater.
C3.	☐	☐	The alpha error rate is changed from .01 to .05.
C4.	☐	☐	The sample size is changed from 30 to 40.

More challenging questions:

	T	F	
C5.	☐	☐	Power is always greater when the effect size is greater.
C6.	☐	☐	Power is always greater when the Type II error (i.e., beta error) is smaller.
C7.	☐	☐	To compute power, all I need to know is effect size, sample size, and alpha.

C8. Which of the following situations would yield the greatest power (assuming alpha and sample size are held constant)?

☐ Null mean = 500, Alternative mean = 510, Standard Deviation = 40

☐ Null mean = 500, Alternative mean = 540, Standard Deviation = 160

☐ Null mean = 500, Alternative mean = 520, Standard Deviation = 60

C9. Consider the shape of the sampling distributions for samples of size n = 4, n = 25, and n = 100. What happens to the sampling distribution of the sample mean when n is increased (assuming nothing else changes)?

☐ Sampling distribution becomes more disperse.

☐ Sampling distribution becomes less disperse.

☐ Sampling distribution remains the same.

C10. So far you have examined the effect of magnitude of difference between the null mean and the alternative mean, standard deviation, sample size, and alpha level on power. Which of the answers below best summarizes the effect of each on power?

☐ More power = large magnitude of difference, larger standard deviation, larger sample, larger alpha.

☐ More power = large magnitude of difference, smaller standard deviation, larger sample, smaller alpha.

☐ More power = large magnitude of difference, smaller standard deviation, larger sample, larger alpha.

☐ More power = smaller magnitude of difference, smaller standard deviation, larger sample, smaller alpha.