BAU
BAHÇEŞEHİR ÜNİVERSİTESİ

BME 3005
Biostatistics

Lecture 7: *Power and Sample Size*

Burcu Tunç Çamlıbel

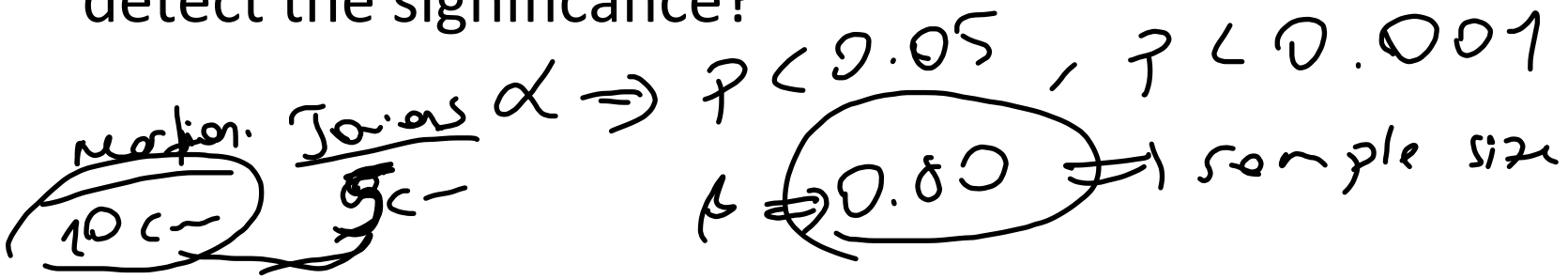
Introduction

- So far, we have learned
  - how to summarize the data
  - ANOVA
  - t-test
  - multiple comparison tests
  - rates and proportions
    - z test
    - contingency tables and chi-square analysis
    - relative risk and odds ratio

What does not significant mean?

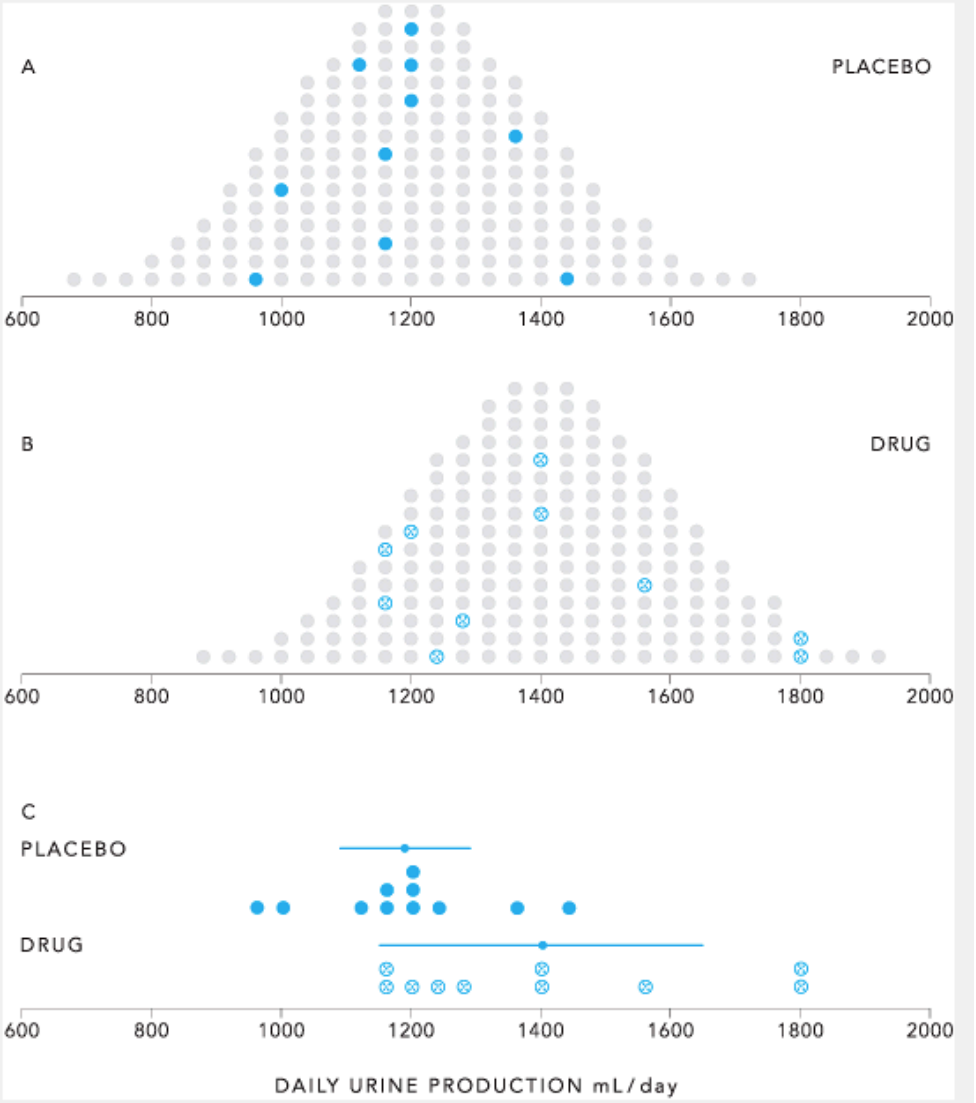
- The researchers conduct an experiment, collect some data that would represent a sample of a population, run a statistical test and report significant or not.

- The main dilemma in the case of not significant is:
  - really not significant?
  - or the statistical procedure lacked the power to detect the significance?



BAU
BAHÇEŞEHİR ÜNİVERSİTESİ

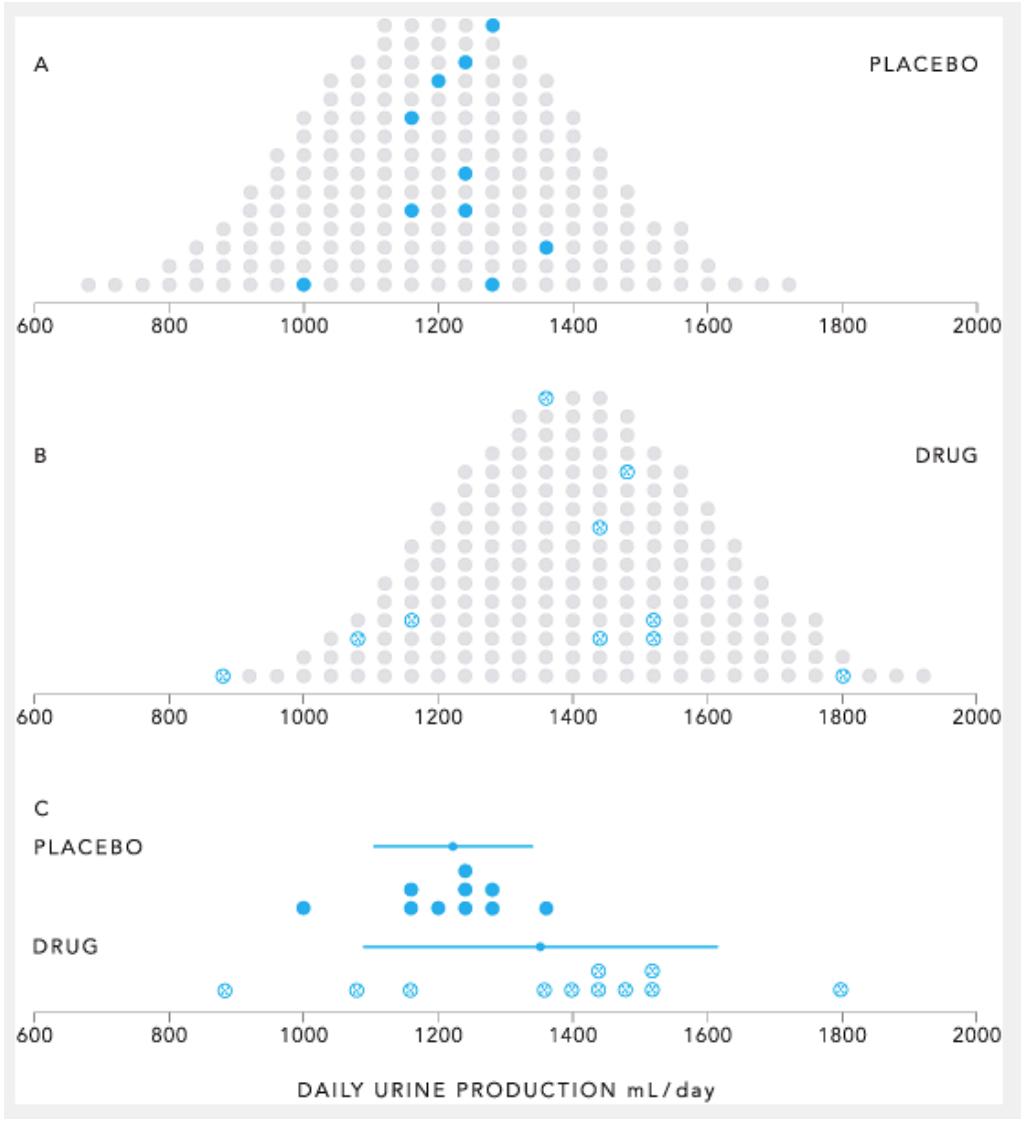
Fig 6.1



Daily urine production in a population of 200 people then they take placebo vs a diuretic drug and the drug has an effect of increasing urine production by 200 ml/day

10 people randomly selected from each group
t = 2.447
2(10-1)=18 degrees of freedom cutoff 2.101
$p<0.05$

Fig 6.2



- Two more random samples of size 10
- This time t=1.71 - less than the cutoff value
- p>0.05 - we would conclude no effect and we would be wrong.

False positive error - Type I error (α)

- In the earlier chapter we were concerned of rejecting the null hypothesis when it was true.
- You conclude that the treatment has an effect, whereas it actually does not have an effect. (False positive)
- We have controlled the likelihood of making a false positive error.
  - P value: I expect to be wrong only 5% of the time when we report that the drug has an effect, when I say that p=0.05.
- This was a false positive error, or type I error (α).

False positive error - Type II error (β)

- Now we are concerned about not rejecting the null hypothesis when it is not true.
- So the drug has an effect, and you fail to report it based on the data sample you have.
- This is a type II error (β), or a false negative error.

Types of Statistical Errors

Table 6–1 Types of Erroneous Conclusions in Statistical Hypothesis Testing

	Actual situation	
Conclude from observations	Treatment has an effect	Treatment has no effect
Treatment has an effect	True positive Correct conclusion 1 - β	False-positive Type I error α
Treatment has no effect	False-negative Type II error β	True negative Correct conclusion 1 - α

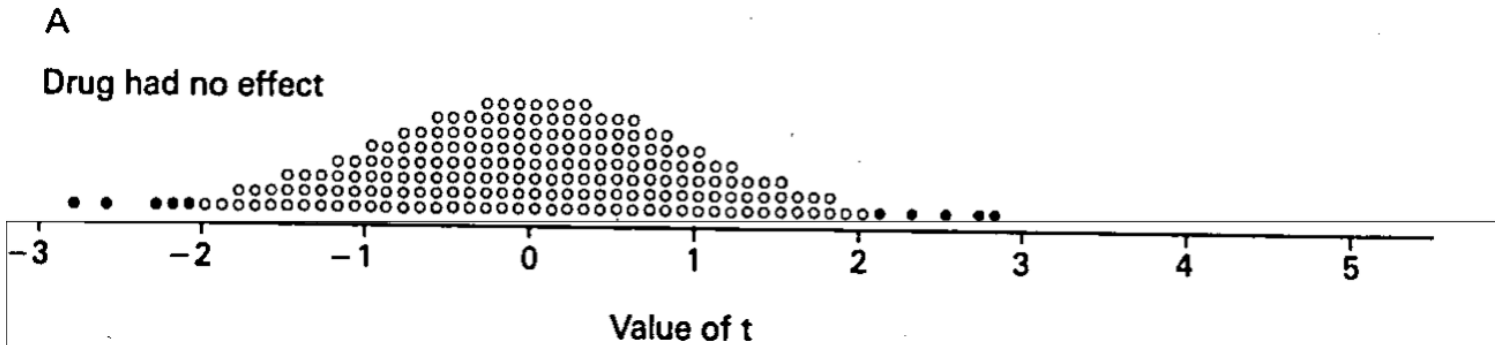
- Positive: Treatment has an effect.
  - You fail to detect it - False negative – β
  - You detect it - True positive – 1-β
- Negative: Treatment does not have an effect.
  - You fail to detect it - False positive – α
  - You detect it - True negative – 1-α

What is power?

- The power of a test is the chance of reporting a true positive.
  - Reporting the treatment has an effect when it actually does have an effect.
  - It is equal to 1-β .
- If a test has power of .55, then there is a 55 percent chance of reporting a statistically significant effect when one is really present.

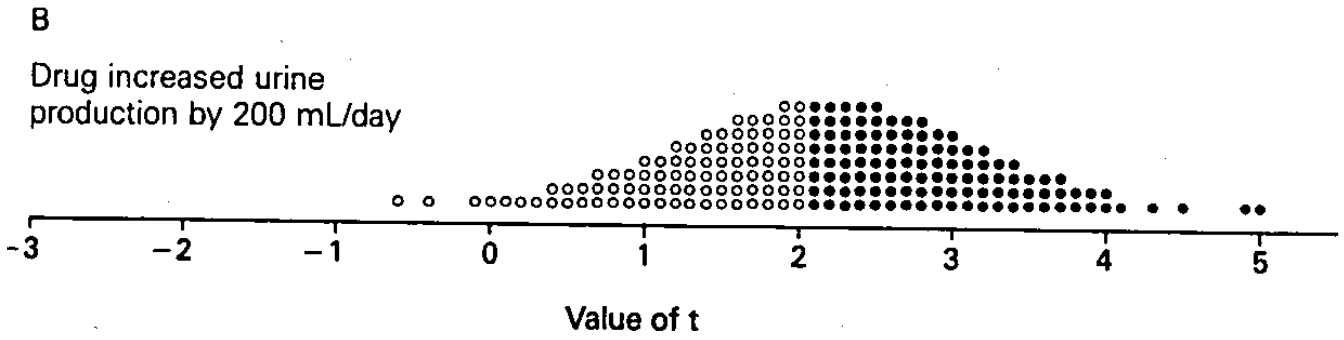
0.80
by complete

Visualization of power



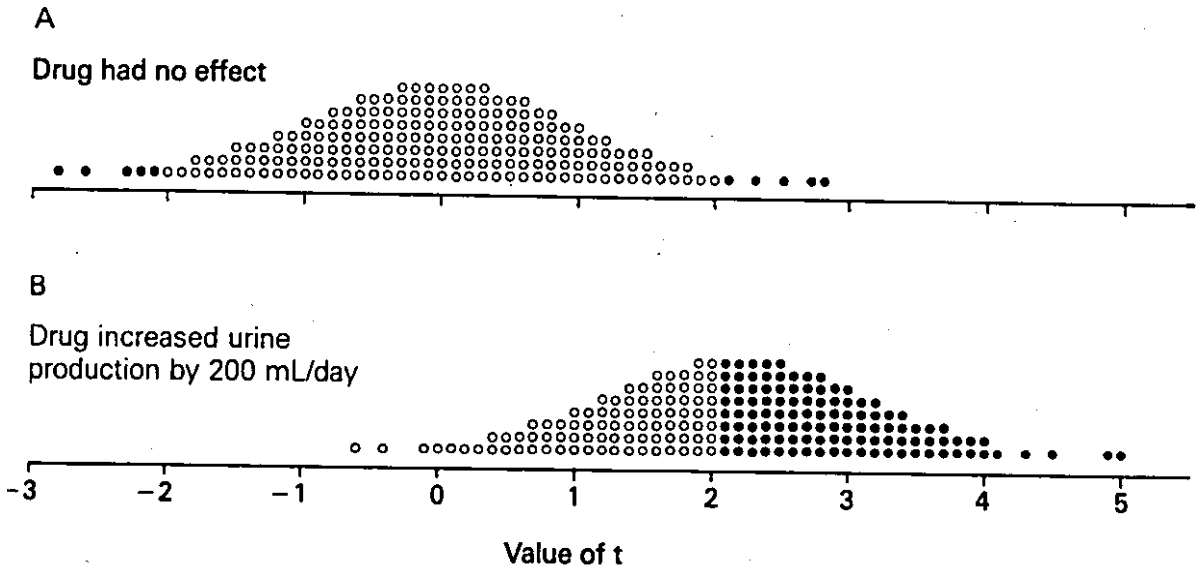
- A- The distribution of t values computed from 200 experiments of sample size = 10 when the drug has no effect
- The cutoff values in the statistical tables are based on the values that one can get when the null hypothesis is true.
- When there is no effect, t values will be centered on “0” with an expected standard deviation.
- The highest 5% of the t values will be contained after the p=0.05 cutoff value.

Visualization of power



- B- The distribution of t values of 200 experiments of sample size = 10 when the drug has an effect
- The t values would be centered at a t>0
- In this example 111/200 falls above a t=2.101 (p=0.05 cutoff value for 2(10-1)=18 degrees of freedom
- There is a 111/200 = 55% chance of detecting the diuretic had an effect.

Visualization of power

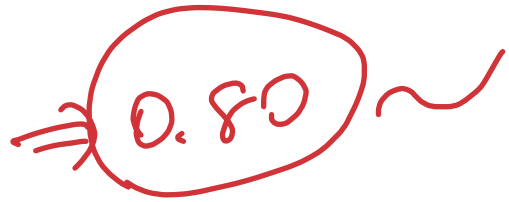


- The power quantifies the chance of detecting a real difference for a given sample size.
- In this case, power is 0.55 (111 out of 200 experiments detected the real effect)
- There is a 45% chance (89/200) of accepting the null hypothesis and being wrong.

What determines a test's power?

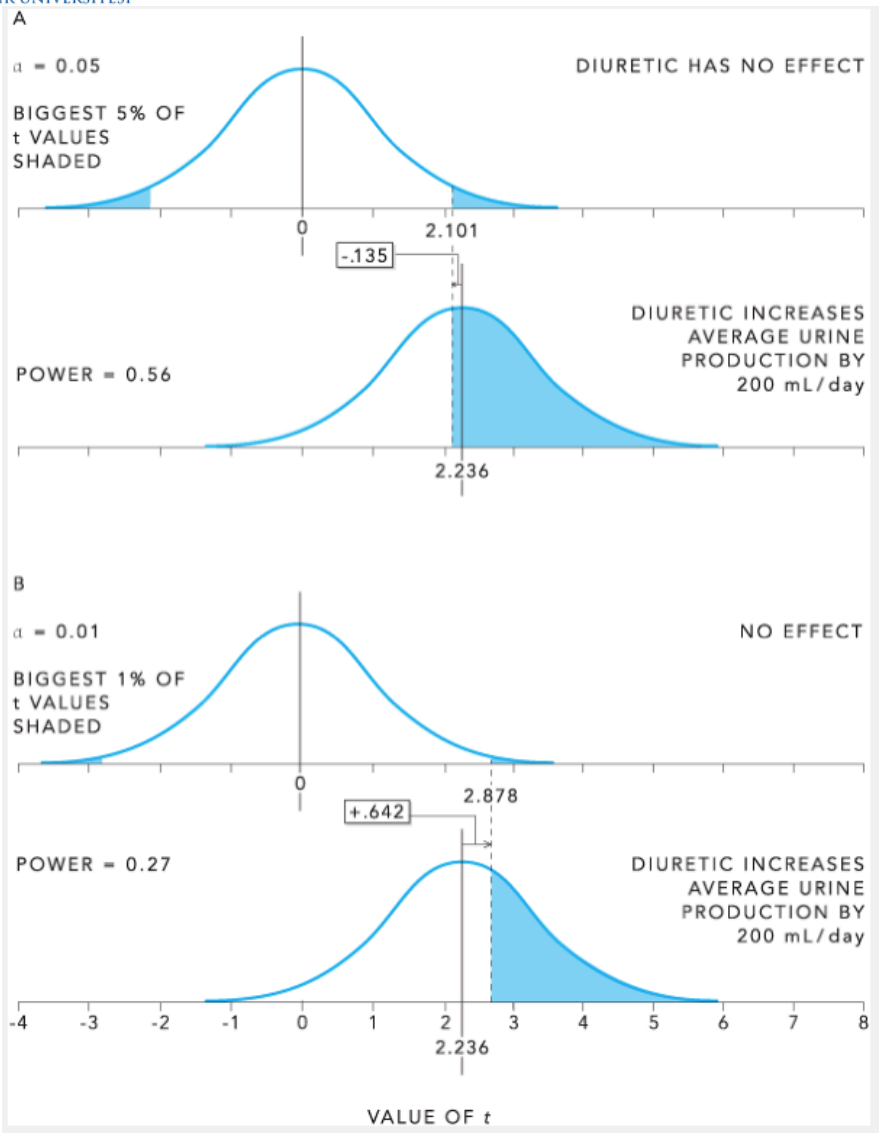
- We want the power of a test as high as possible.
  – reducing the type II error
- The size of the treatment effect influences how easy it is to detect it.
  – Larger the effect, easier to detect it.
- Type I and Type II errors are intertwined.
  – Whenever you would like to have stronger evidence to detect a significance, you reduce the power, or increase the likelihood of missing a true effect.
  – reducing increases , or false negative.
- The only way both and can be reduced is by increasing the sample size and detecting the difference more confidently if there is one.

The power depends on



- The risk of error you will tolerate when rejecting the null hypothesis
- The size of the difference you wish to detect relative to the amount of variability in the population
- The sample size

The size of Type I error α



The distribution of all t values for all possible experiments with a sample size = 10 out of 200 people

For $\alpha = 0.05$ $t = 2.101$

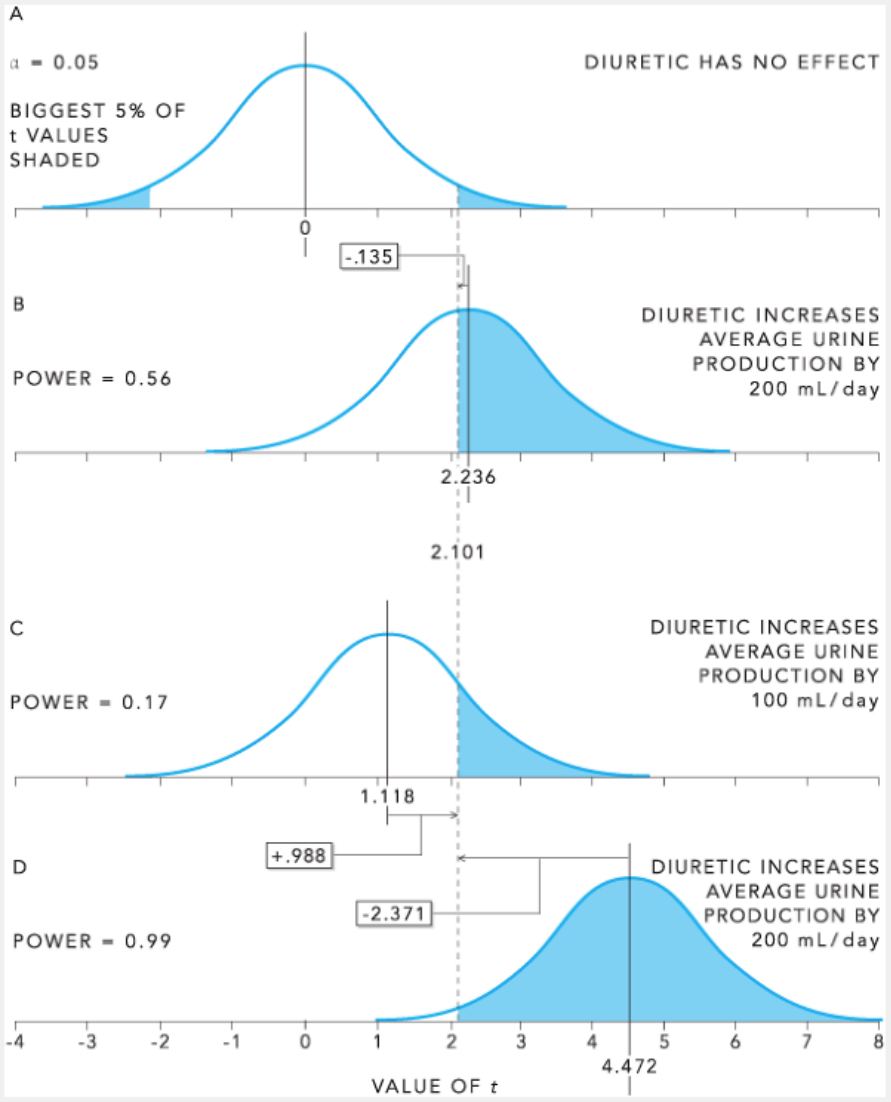
$1-\beta = 0.55$

For $\alpha = 0.01$ $t = 2.878$

$1-\beta = 0.45$

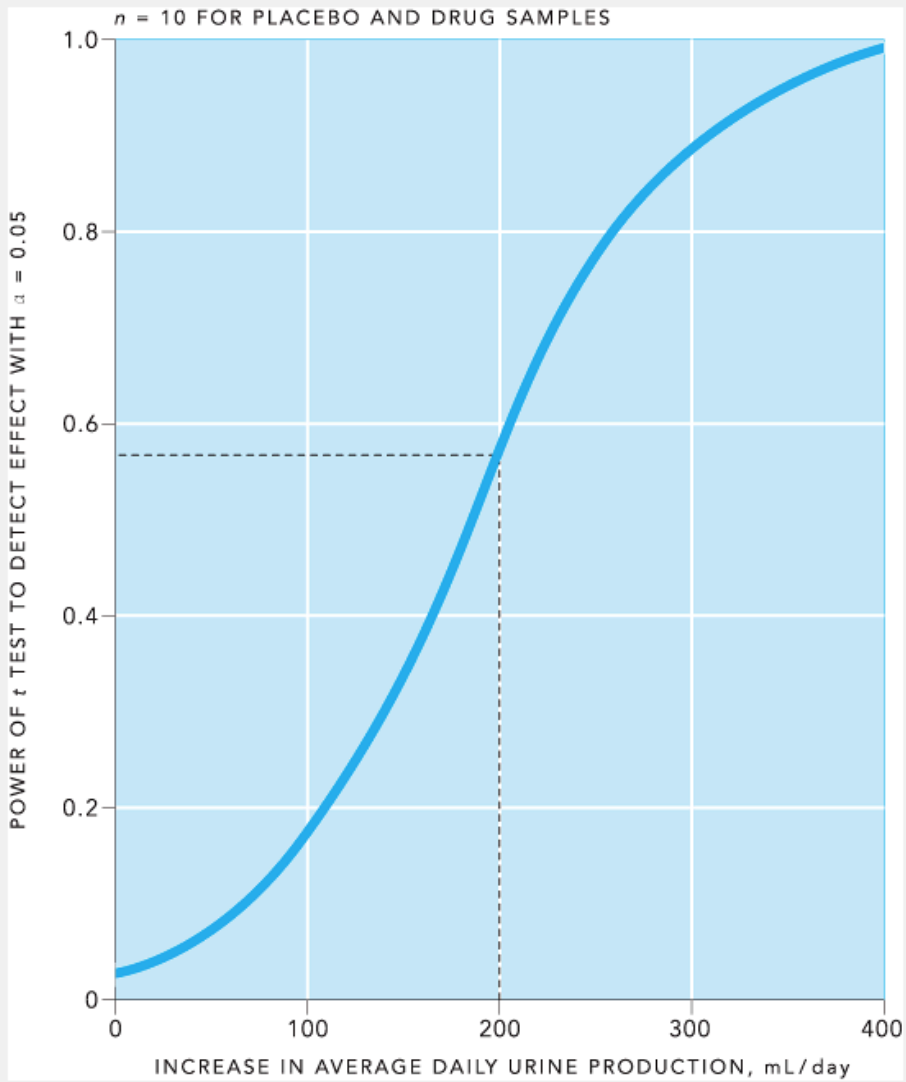
So, when we would like to increase the confidence of detecting a real difference, we decrease the power or the chance of detecting a true positive.

The size of the Treatment Effect



- As the size of the treatment effect increase, the numerator of the t- test will be bigger.
- The distribution of t-values will be centered at a higher value.
- The power or the chance of detecting a true positive will be bigger.

Power Chart for Increase of Effect



The power of a t-test to detect a change in urine output based on experiments with two groups of people, each containing 10 individuals.

The t-test has a power of .55 of detecting a 200 ml/day change.

The power increases as the actual drug effect gets larger.

What is the power at 0? and why?

The Population Variability

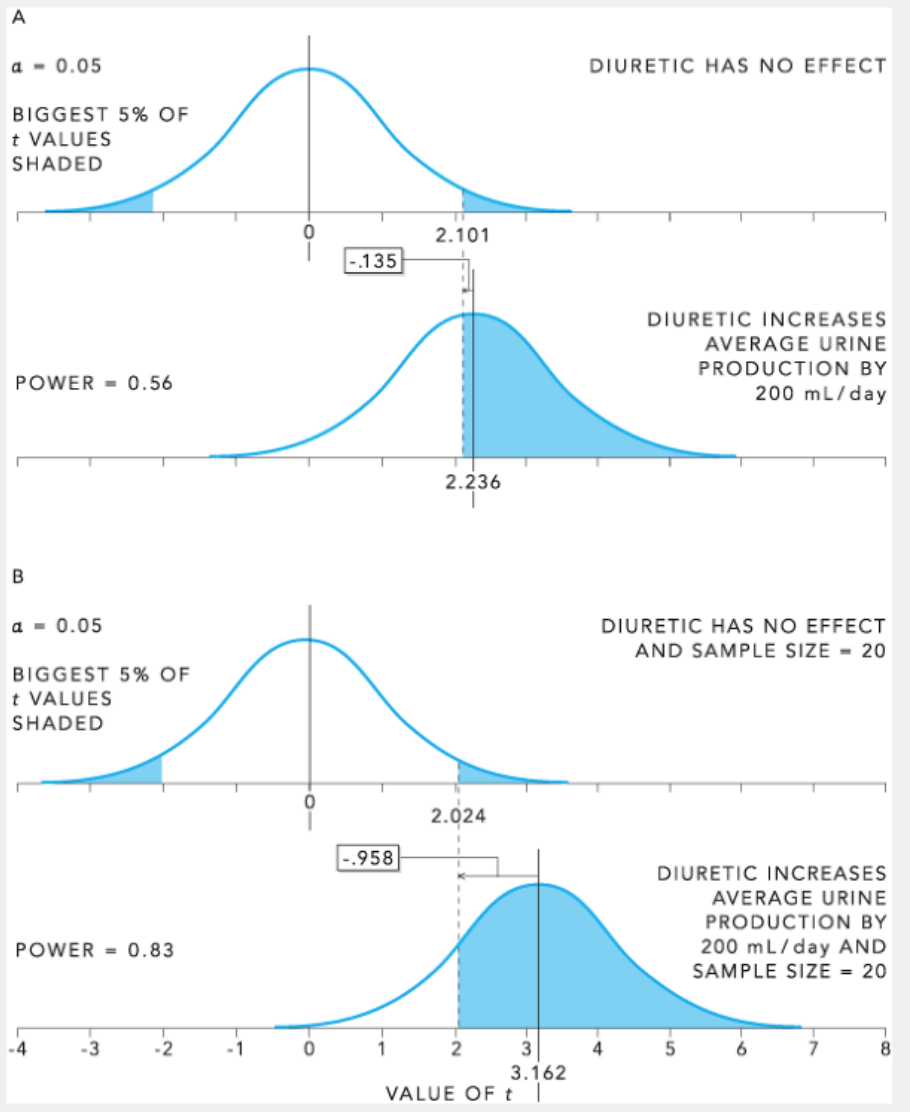
$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(s^2/n_1) + (s^2/n_2)}} \quad \xrightarrow{\text{if n1=n2}} \quad t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(s^2/n_1) + (s^2/n_2)}}$$

$$t' = \frac{\delta/\sigma}{\sqrt{2/n}} = \frac{\delta}{\sigma}\sqrt{\frac{n}{2}}$$

δ: the change in the population mean value after the treatment

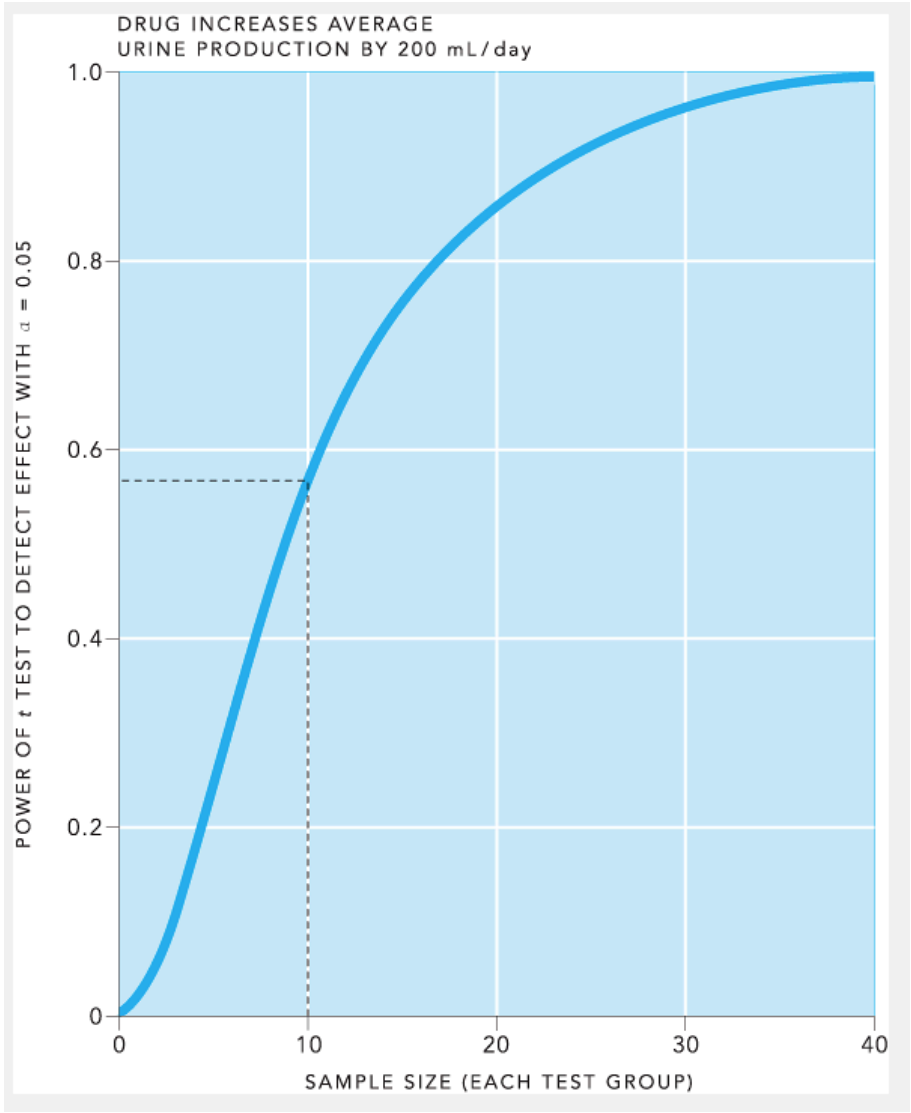
- t depends on the change in the mean response normalized by the population standard deviation.
- As the variability in the population, ,decreases, the power of the test increases.
- Φ=δ/σ is called noncentrality parameter.

Sample Size



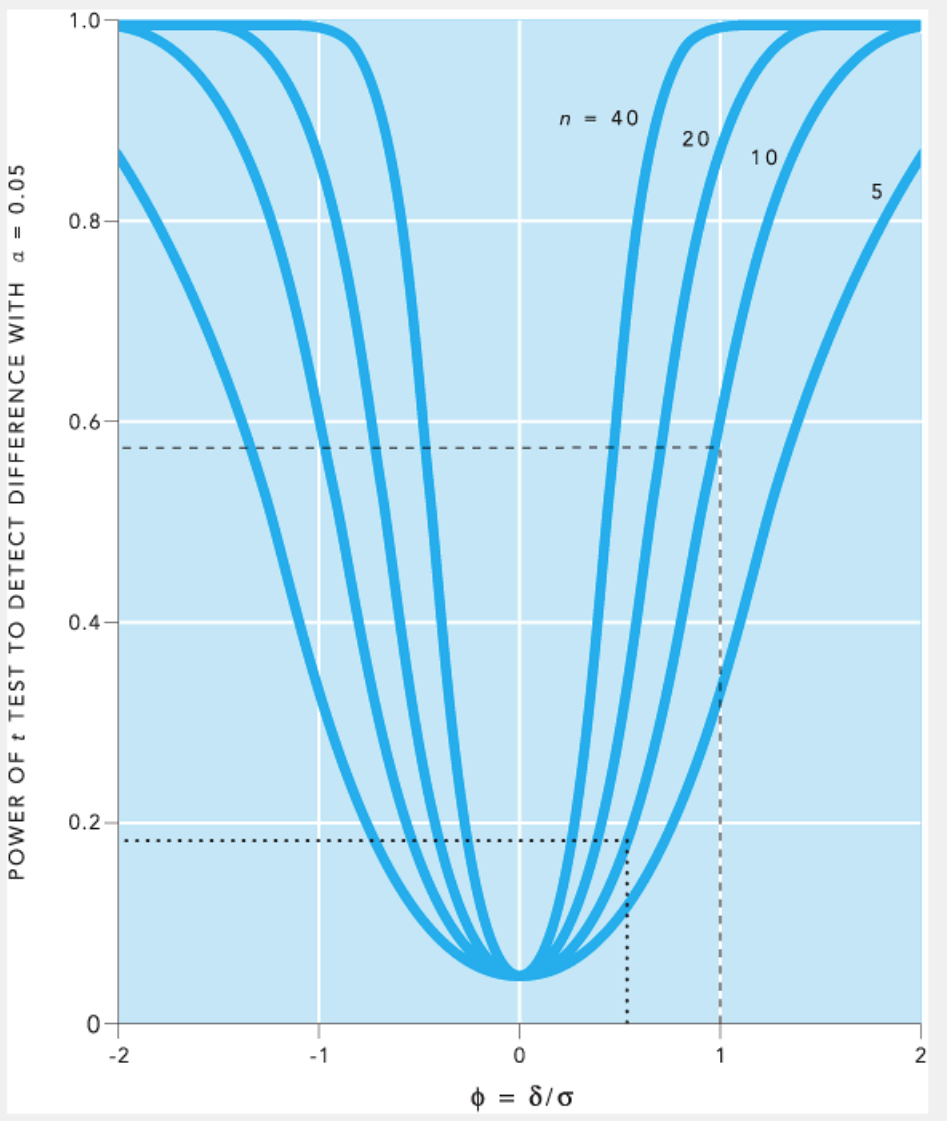
- In most of the cases, researchers cannot control α or Φ=δ/σ, but they can increase sample size to increase the power.
- Increasing sample size increases the degrees of freedom and result in a decrease in the cutoff t value.
- The value of t also increases as the value n increases.

Power Chart for Increase of Sample Size



- The effect of sample size on power of t test for a 200 ml/day increase with 200 ml/day standard deviation
- Estimating the sample size required to detect a clinically significant difference is the major practical use of the power computations.

The power function for t-test for = 0.05



- δ: size of the change we wish to detect
- σ: standard deviation
- n: sample size

In the case of unequal sample size, use the size of the smaller sample in the power analysis.

Example

Table 4–2 Comparison of Anesthetic Effects on the Cardiovascular System

	Halothane (n = 9)		Morphine (n = 16)	
	Mean	SD	Mean	SD
Best cardiac index, induction to bypass, $L/m^2 \cdot min$	2.08	1.05	1.75	.88
Mean arterial blood pressure at time of best cardiac index, mmHg	76.8	13.8	91.4	19.6
Total peripheral resistance associated with best cardiac index, $dyn \cdot s/cm^5$	2210	1200	2830	1130

- Cardiac index = rate at which heart pumps the blood/body surface area

$$s^2 = \frac{(9-1)(1.05^2) + (16-1)(.88^2)}{9+16-2} = 0.89$$

$$t = \frac{2.08 - 1.75}{\sqrt{(.89/9) + (.89/16)}} = 0.84$$

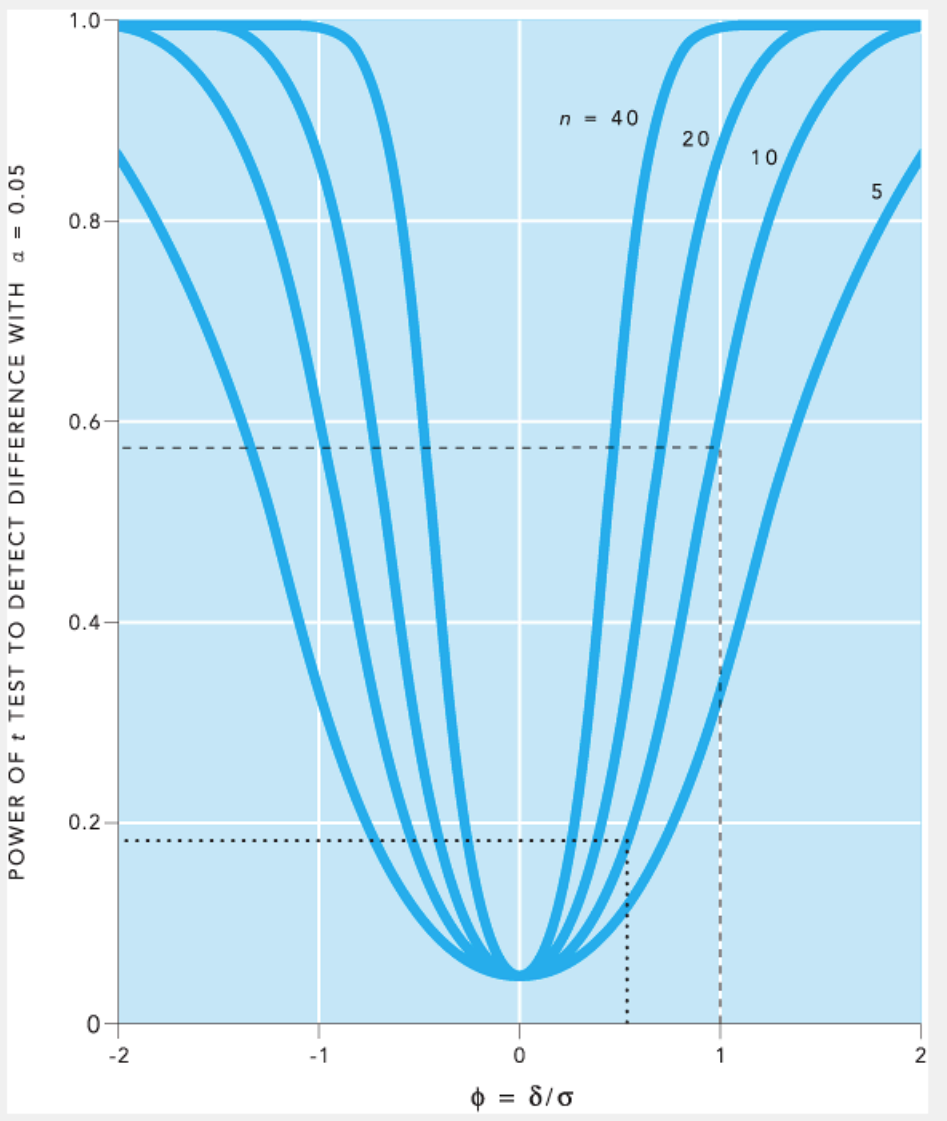
v=9+16-2 = 23, t cutoff = 2.069
Halothane and morphine don't differ in their effect on cardiac index.

The difference was (2.08-1.75)/2.08*100 = 15% change

Example (cont)

- Question: *What is the power of this experiment to detect a 25% change in cardiac index?*
- 25% of 2.08 = 0.52
- pooled variance = 0.89. Compute the standard deviation as, sqrt(variance) = 0.94.
- $\Phi = \delta/\sigma = 0.52/0.94 = 0.553$
- Sample size = 9 (the size of the smaller group)
- Check Fig 6.9 with the given parameters

Example (cont)



Φ= 0.553

n=9

power = 0.16!

It is very unlikely that this experiment would be able to detect a 25% change in cardiac index.

BAU
BAHÇEŞEHİR ÜNİVERSİTESİ

Rules of Power

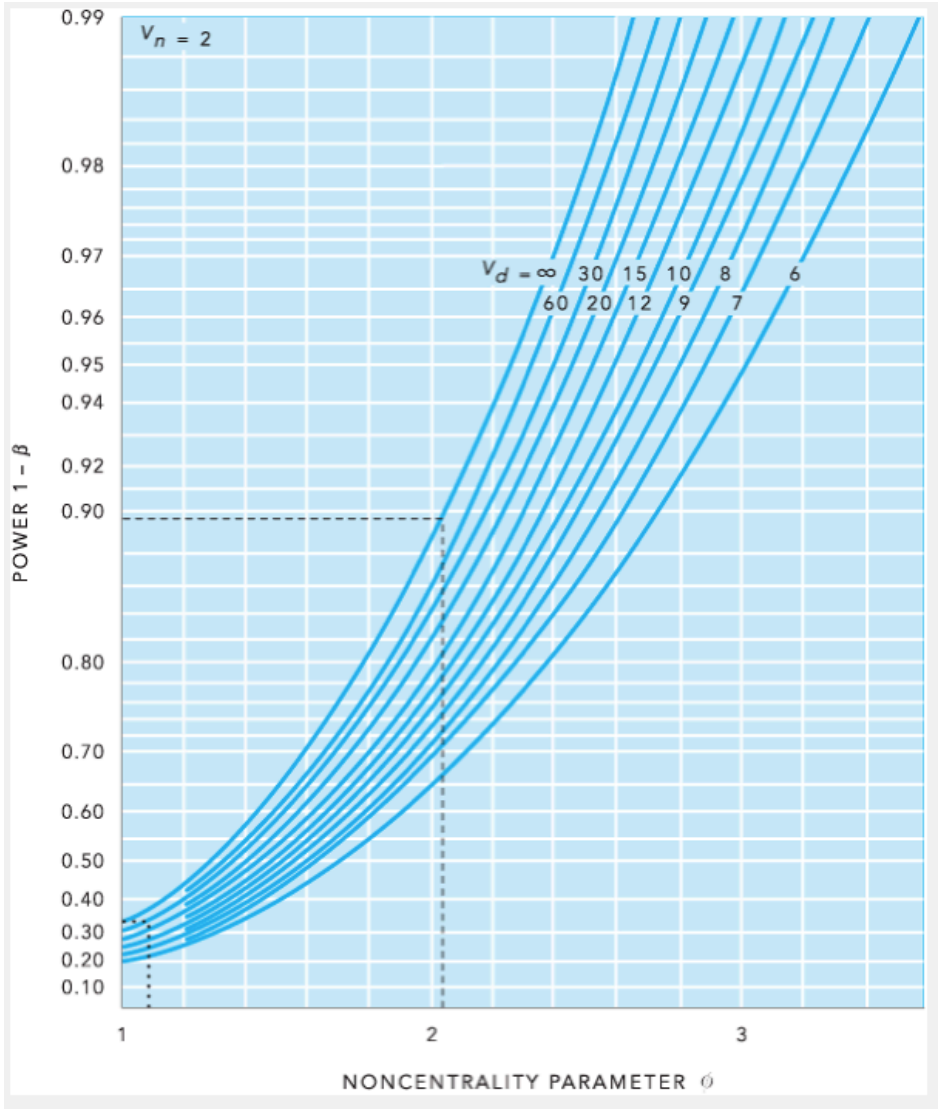
- The power of a test is 1-β and tells us the likelihood of null hypothesis getting rejected when the treatment has an effect. (True positive)
- The lower the α, or the lower the error rate of producing a false positive, the lower the power of the test.
- Smaller sample size results in lower power.
- Smaller size of the treatment effect wrt. the population σ, the lower the power.
- The exact procedure to compute the power of a test depends on the test itself.

Power and Sample Size for ANOVA

$$\phi = \frac{\delta}{\sigma}\sqrt{\frac{n}{2k}}$$

- δ: the minimum difference we would like to detect
- σ: the standard deviation
- n: sample size of each treatment group
- k: number of treatment groups

Power Chart for ANOVA

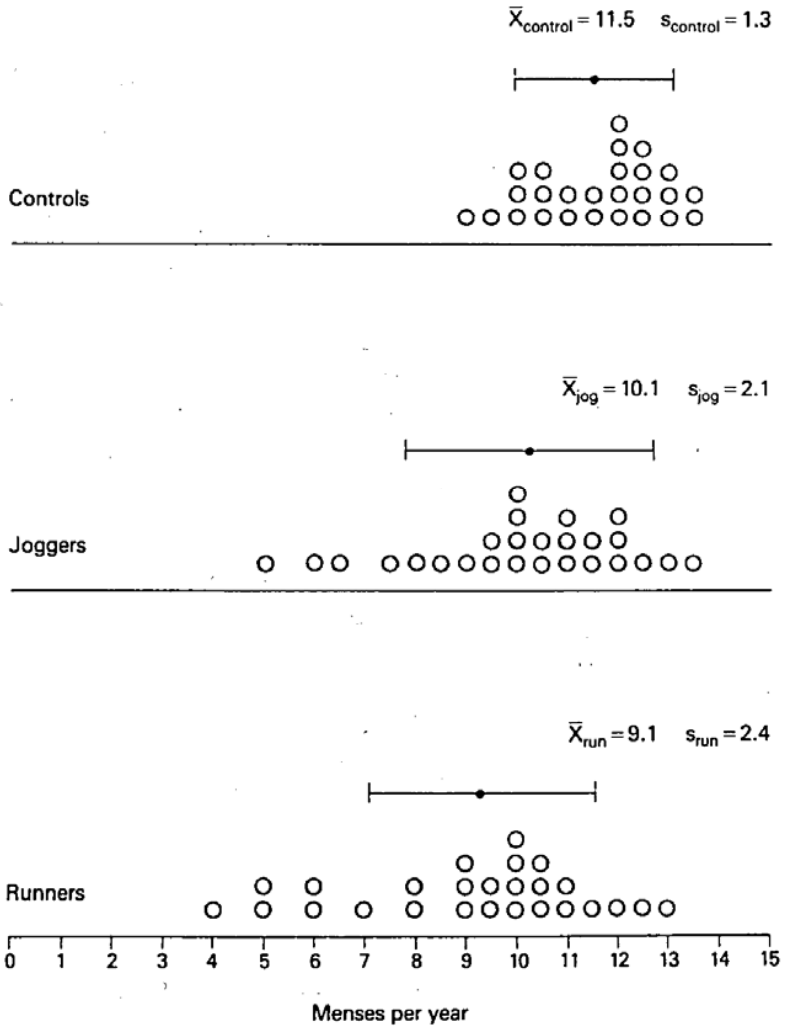


- $v_n = k-1$
- $v_d = k(n-1)$

- This chart is for $v_n = 2$ and $\alpha = 0.05$ as the denominator degrees of freedom changes.

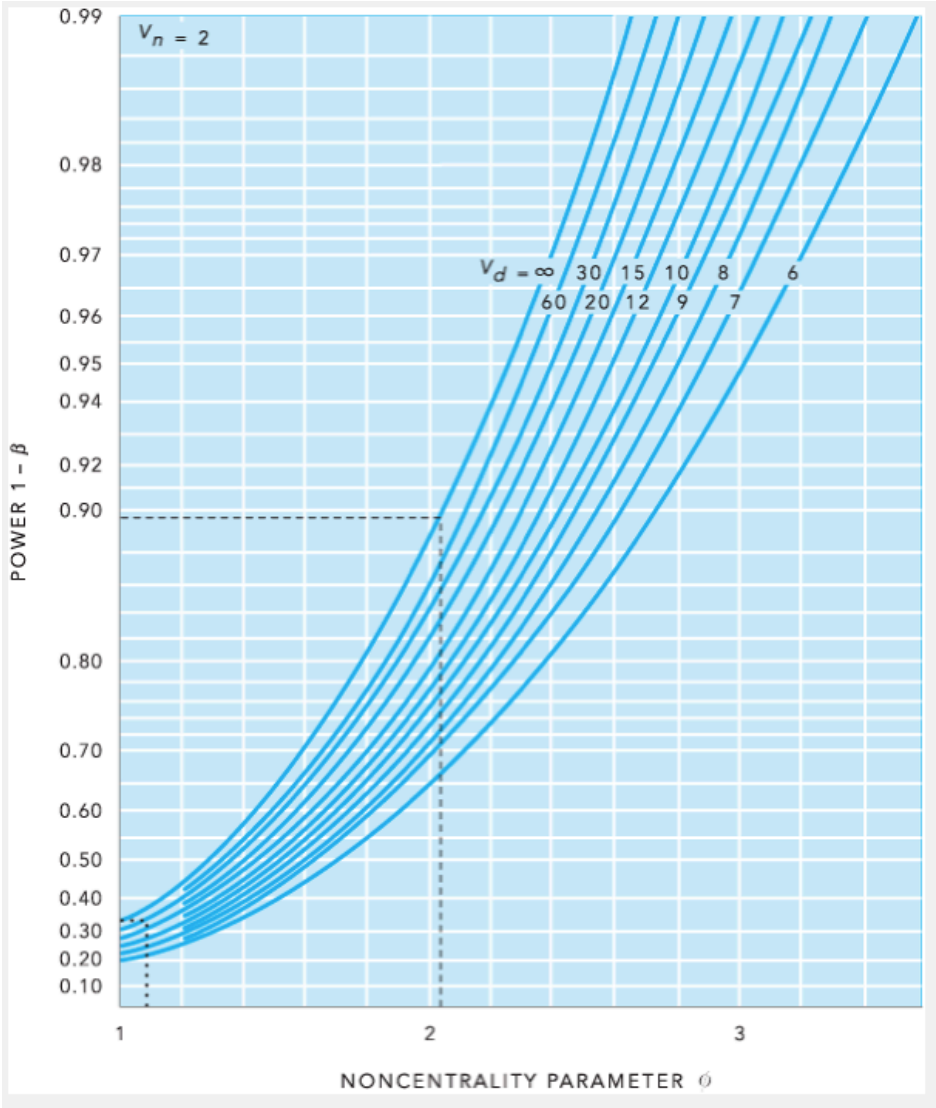
- More power charts for ANOVA in Appendix B

Example



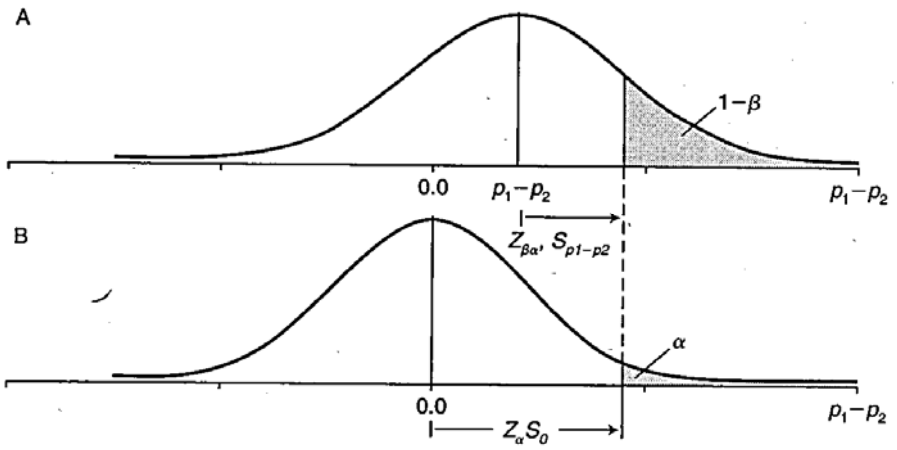
- Effect of running on # menses per year
- Suppose we want to detect a change of
  - $\delta = 1$ menses/year
  - with $\sigma = 2$ menses/year
  - among $k = 3$ groups of women
  - $n = 26$ women in each group
  - with 95% confidence ($\alpha = 0.05$)

Example (cont)



- v_n =2, v_d =3(26-1)=75
- Power is .32!
- If we want to increase the power, we need to increase the sample size.
  - if we increase n to 100-- Φ= 2.04, v_d = 3(100- 1)=297, power = 0.90
  - if we increase n to75-- Φ= 1.77, v_d = 3(75-1)=222, power = 0.80

Power and Sample Size for Comparing Two Proportions



$$z = \frac{\hat{p}_1 - \hat{p}_2}{s_{p_1-p_2}}$$

$$s_{p_1-p_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

- The distribution of all possible values of differences between two proportions follows a normal distribution with mean p1-p2 and standard deviation s_{p1-p2}.

Power for Comparing Two Proportions

Table 6-2 Percentile Points of the Standard Normal Distribution (One Tail)

Fraction of distribution *below z* (β)	Fraction of distribution *above z* (1 − β) Power	*z*
.001	.999	−3.0902
.005	.995	−2.5758
.010	.990	−2.3264
.020	.980	−2.0538
.050	.950	−1.6449
.100	.900	−1.2816
.200	.800	−0.8416
.300	.700	−0.5244
.400	.600	−0.2534
.500	.500	0.0000
.600	.400	0.2534
.700	.300	0.5244
.800	.200	0.8416
.900	.100	1.2816
.950	.050	1.6449
.980	.020	2.0538
.990	.010	2.3264
.995	.005	2.5758
.999	.001	3.0902

- z is the lower tail value that defines the lower 100*β percent of the population.
- When z = -0.842, the power will be 80 percent and 20 percent of the distribution will be below this cutoff.

Power of z

- To achieve a power of 100(1-β) percent, 100(1-β) percent of this distribution should be above (p1- p2)+$z_\beta s_{p1-p2}$
- z_β is the lower tail value on the standard normal distribution that defines the lowest 100β percent of the distribution. $z_\alpha s_0$ is the cutoff value in the normal distribution of null hypothesis.

$$z_{\beta(1)} = \frac{z_\alpha s_0 - (p_1 - p_2)}{s_{p_1 - p_2}}$$

Example

- 13.1% of 61 patients given halothane, and 14.9% of 67 patients given morphine died.
- What is the power of this study to detect a 30% difference in mortality from 14 to 10 percent with 95% confidence?
- p1=.14, p2=.10

$$s_{p_1-p_2} = \sqrt{\frac{.14\,(1-.14)}{61} + \frac{.10\,(1-.10)}{67}} = .0576$$

$$\hat{p} = \frac{.14 \cdot 61 + .10 \cdot 67}{61 + 67} = .119 \qquad s_0 = \sqrt{\frac{.119\,(1-.119)}{61} + \frac{.119\,(1-.119)}{67}} = .0573$$

$$z_{\beta(1)} = \frac{1.96 \cdot .0573 - (.14 - .10)}{.0576} = 1.255$$

$z_{0.05} = 1.96$

From Table 6.2, power is 11%

Problems 6.1-6.3, Primer of Biostatistics, Glantz (6th Edition)

6-1 Use the data in Table 4–2 to find the power of a *t* test to detect a 50 percent difference in cardiac index between halothane and morphine anesthesia.

6-2 How large a sample size would be necessary to have an 80 percent chance of detecting a 25 percent difference in cardiac index between halothane and morphine anesthesia?

6-3 Use the data in Table 4–2 to find the power of the experiments reported there to detect a 25 percent change in mean arterial blood pressure and total peripheral resistance.

Problems 6.9&6.10, Primer of Biostatistics, Glantz (6th Edition)

6-9 What is the power of the experiment in Prob. 5–4 to detect a situation in which nefazodone and psychotherapy each causes remission one-third of the time, and nefazodone and psychotherapy cause remission one-half of the time? Assume that the same number of people take each treatment as in Prob. 5–4. Use $\alpha = 0.05$.

6-10 How large would the sample size need to be in Prob. 6–9 to reach 80 percent power?