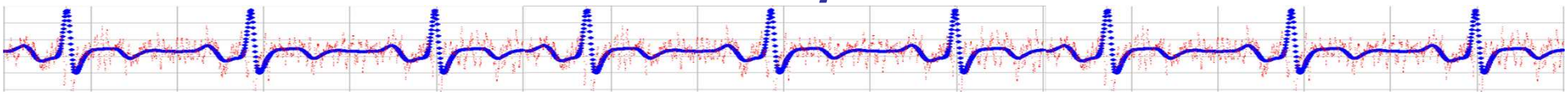
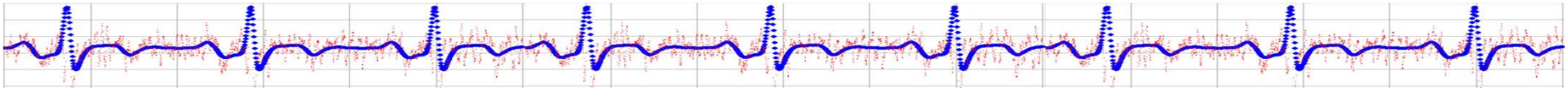
Empirical Research Methods in Information Science

IS 4800 / CS6350



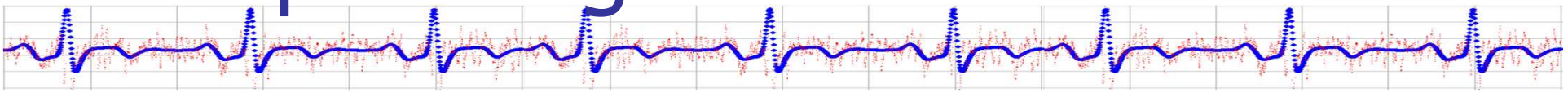
Lecture 14

Outline



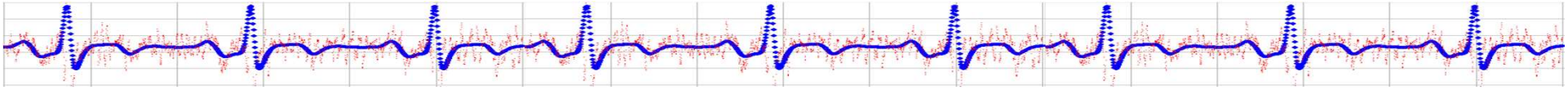
- Reading assessment
- Schedule update
- Finish correlation
- Inferential statistics and power

Upcoming



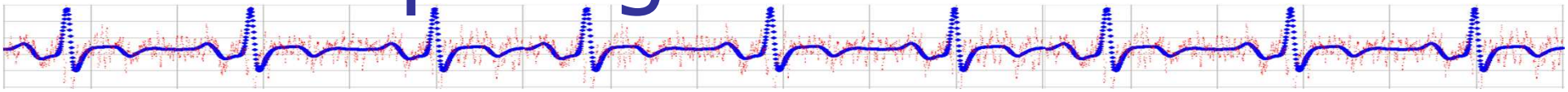
- Assignment I5 due (Friday, EOD)
- Enjoy your spring break!
- Tuesday class – review, team up (T1)
- Wednesday class – midterm exam
- T1 assignment deadline will be adjusted back a bit … updates on Piazza

Pearson's r



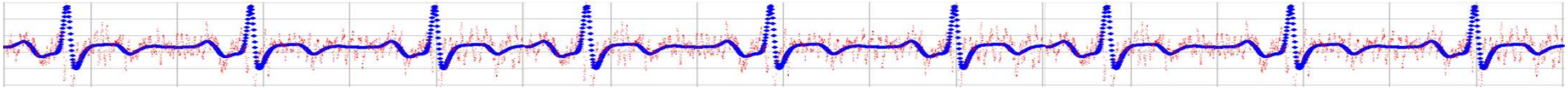
- https://www.socscistatistics.com/tests/pearson/Default2.aspx

Comparing r's



- If you want to make statements about how large one correlation is relative to another.
  - e.g. one is twice as large as another
- Don't compare r's directly…
- Compare r^2 ("proportionate reduction in error")

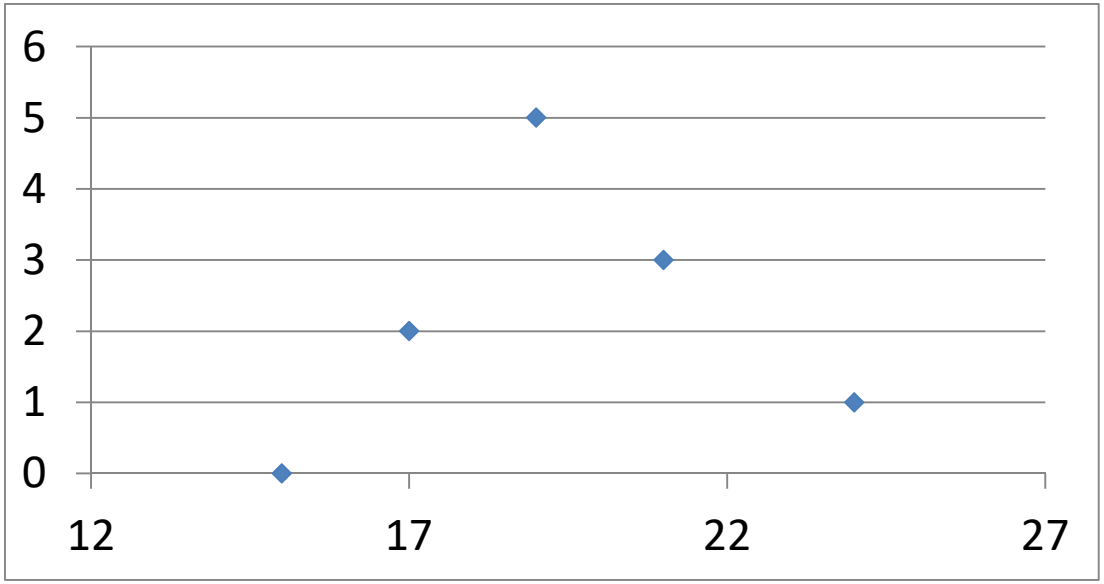
Other measures of association



- Point-biserial [r_{pb}]
  - One numeric & one binary (nominal) measure
  - Just dummy code the nominal (0 and 1) and use Pearson correlation.
- Spearman rank correlation coefficient (Spearman's rho) [ρ or r_s]
  - Two ordinal measures (or for transformed numeric measures if non-linear)
  - Replace each value with its rank order
  - Compute Pearson correlation with ranks
  - Measures degree of monotonicity

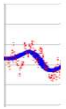
Sample Data for Spearman rho

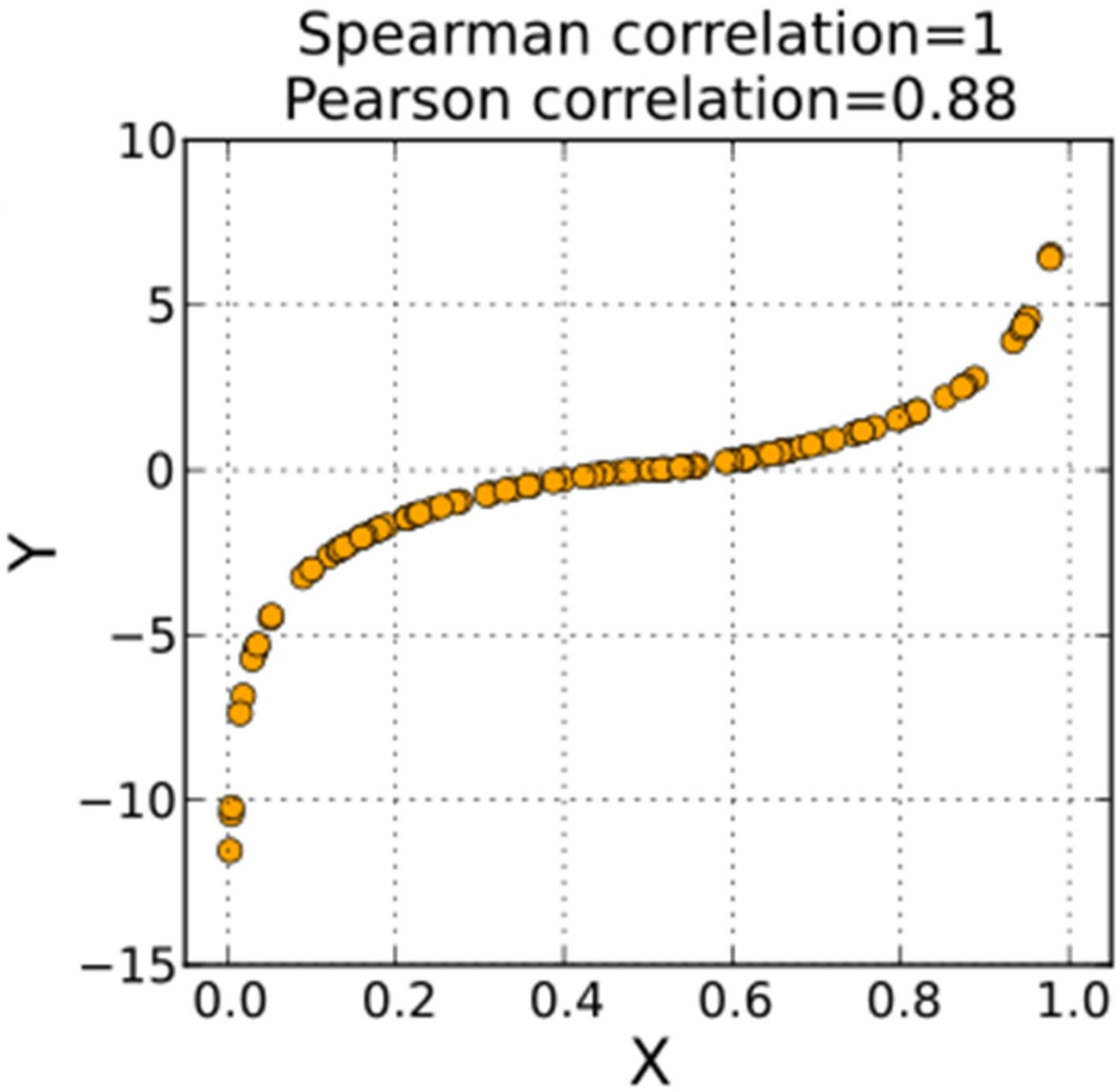
Monitor Size	21	24	17	19	15
Productivity	3	1	2	5	0

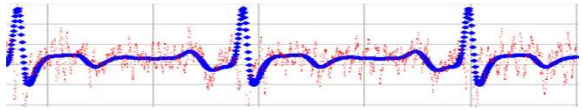


rho = 0.3

*r = 0.18

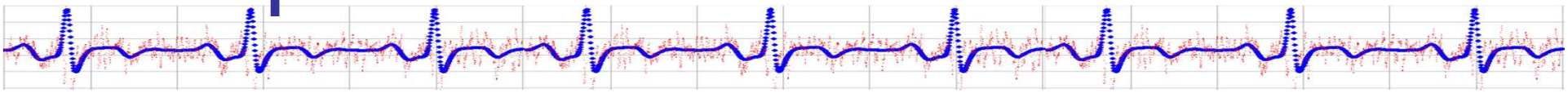






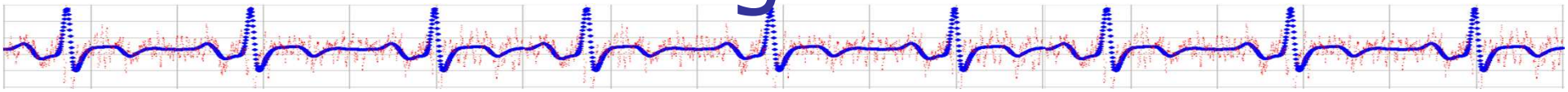
Spearman measures degree of monotonicity.

Spearman's rho



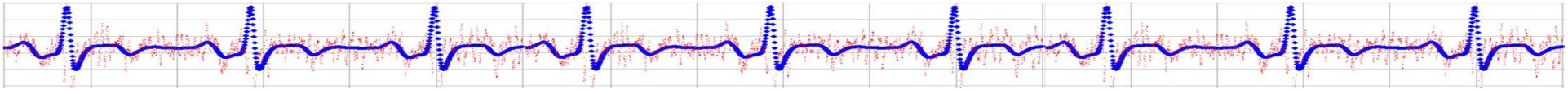
- https://www.socscistatistics.com/tests/spearman/default2.aspx

Two meanings of 'correlation'



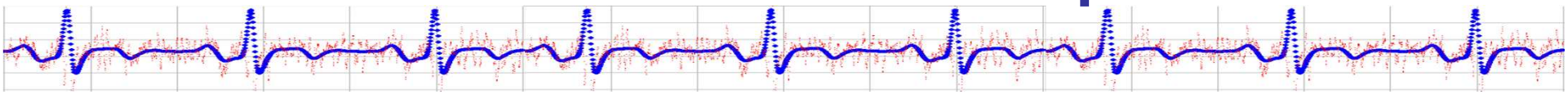
- Correlation statistic vs.
- Correlational research model

Example: Net Latency & Satisfaction

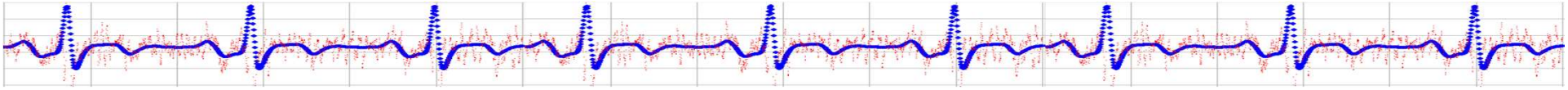


- NU ITS wants to save money by switching to slower wireless routers, and wants to assess the impact this will have on student satisfaction. You want to know how slow things have to get before students start complaining.
- You have the crew implement a program that randomly chooses a network latency (between 0s and 10s) every time a student logs into NUwave, then adds that latency to every network access from them. After 10 minutes of use a web form pops up asking students to rate their degree of satisfaction with NU ITS (10-item).
- What kind of study is this?
- What statistic would you use to evaluate results?
- Assuming r = -0.8, p=.021, what would you conclude?
- Assuming r = 0.1, p=.342, what would you conclude?

What do you do if your data is clearly not unimodal & symmetric OR there is a clear non-linear relationship?

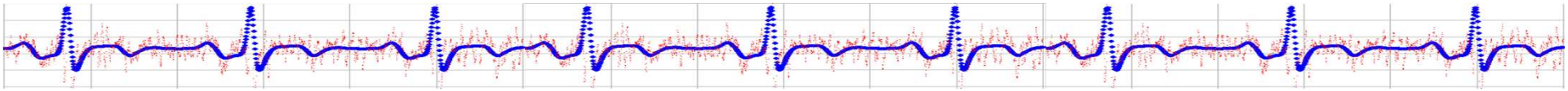


Parametric vs. Non-parametric Statistics

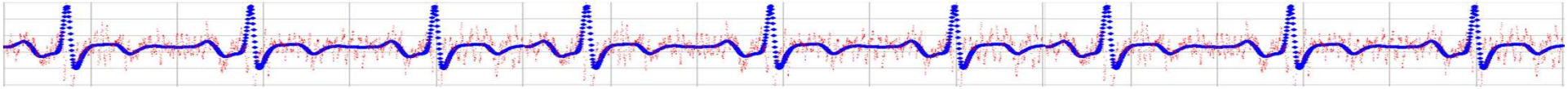


- Non-parametric statistics that do not rely on data belonging to any particular distribution
- E.g., Pearson r is a parametric statistic (assumes underlying distributions are normal – can be described using parameters – mean & stddev)
- E.g., Spearman rho is non-parametric

Intro to Power & Effect Size

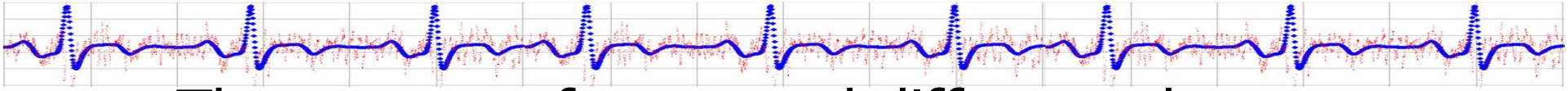


Power



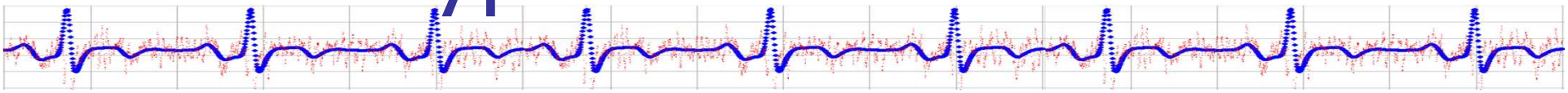
- The "power" of a statistical test is its ability to detect differences in data that are inconsistent with the null hypothesis.
  - p(rejecting H0|H1)
- aka – the ability to find a significant result, if your hypotheses are actually true.
- What is it called when this fails (ie, accepting H0 when H1 is true)?
- Why is this a bad situation?

Effect size



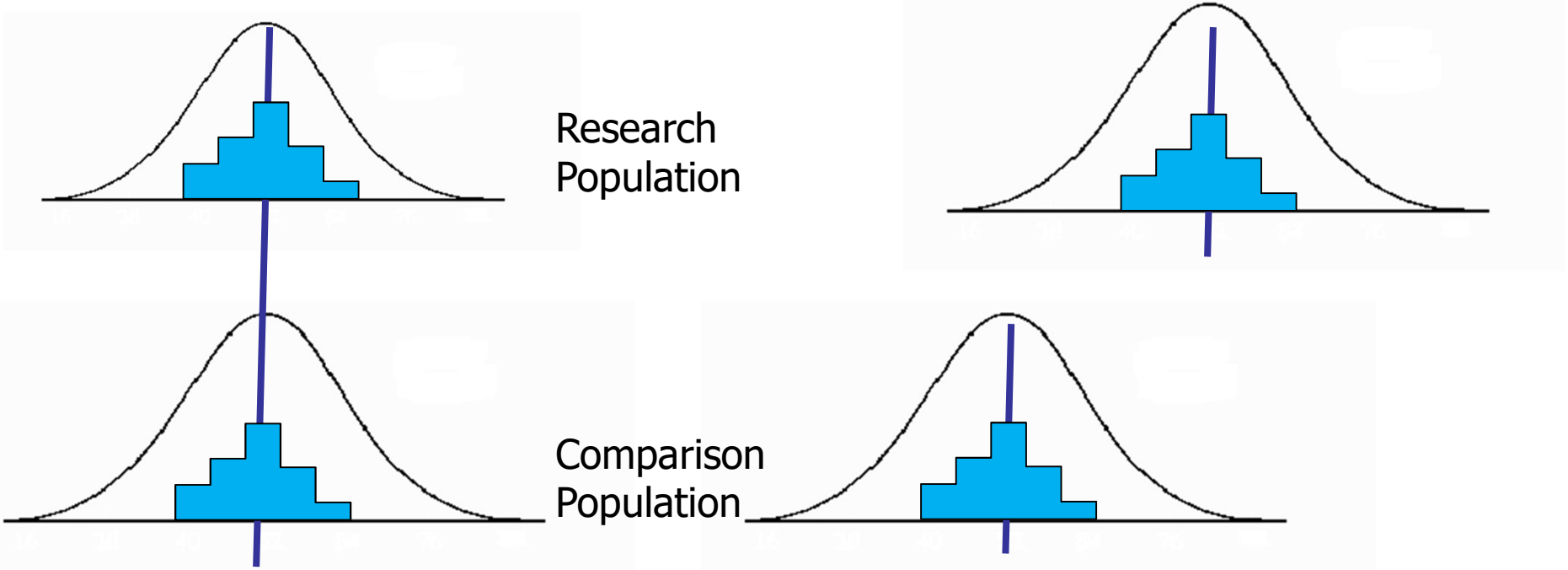
- The *amount* of measured difference between study conditions.
- The greater the effect size, the easier it is to show there is a significant difference in your study (ie, the greater the power).
- Effect size formula is different for each hypothesis test procedure.
- Tabulated standard values for "small", "medium", and "large" effect sizes.
- Only talk about effect size IF significance is established – but then DO present it in your results.

The typical situation

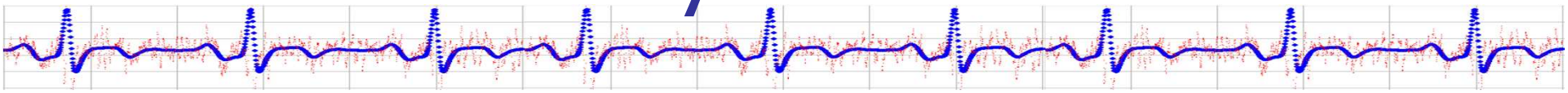


H0 Actually True

H1 Actually True

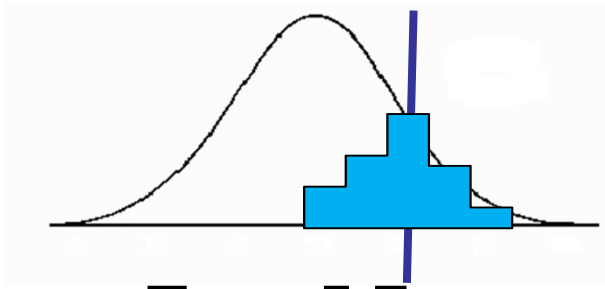


The unlucky situations



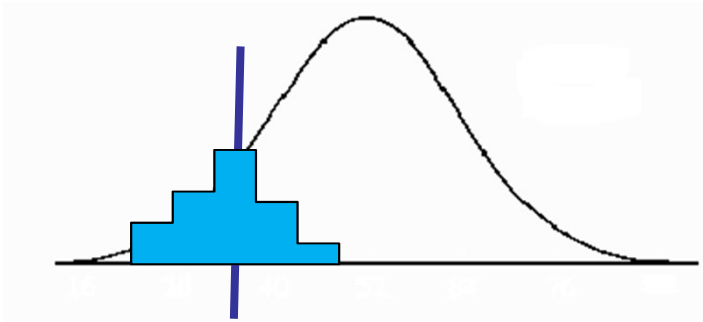
H0 Actually True

H1 Actually True

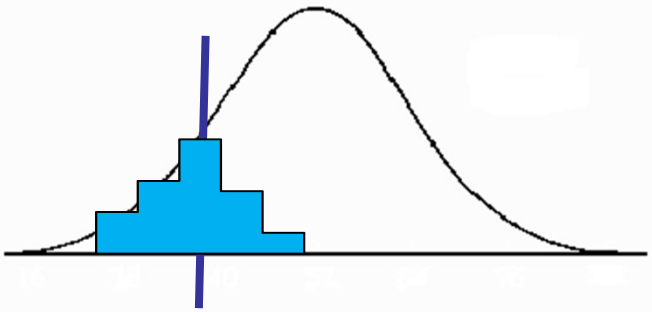


Research Population

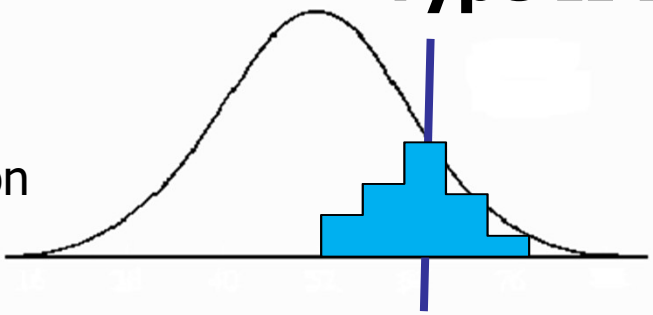
Type I Error



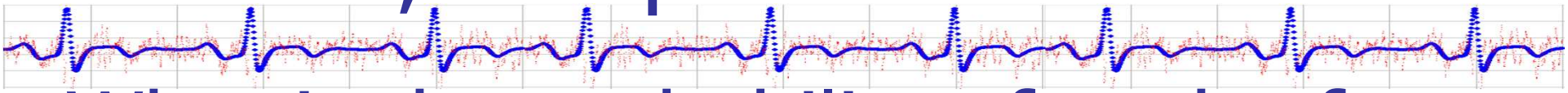
Type II Error



Comparison Population



Relationship between alpha, beta, and power.

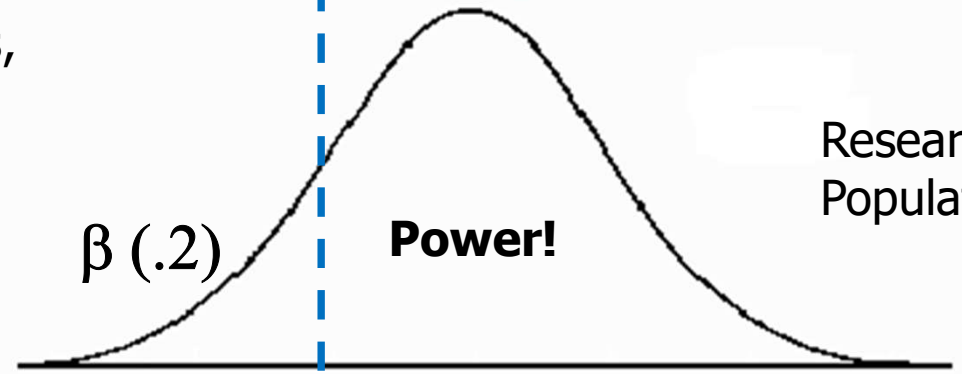


What is the probability of each of these situations occurring?

	"The Truth"	
	H1 True	**H1 False**
Decide to Reject H0 & accept H1	Correct p = power	Type I err p = α
Do not Reject H0 & do not accept H1	Type II err p = β	Correct p = 1-α

Relationship between power and effect size

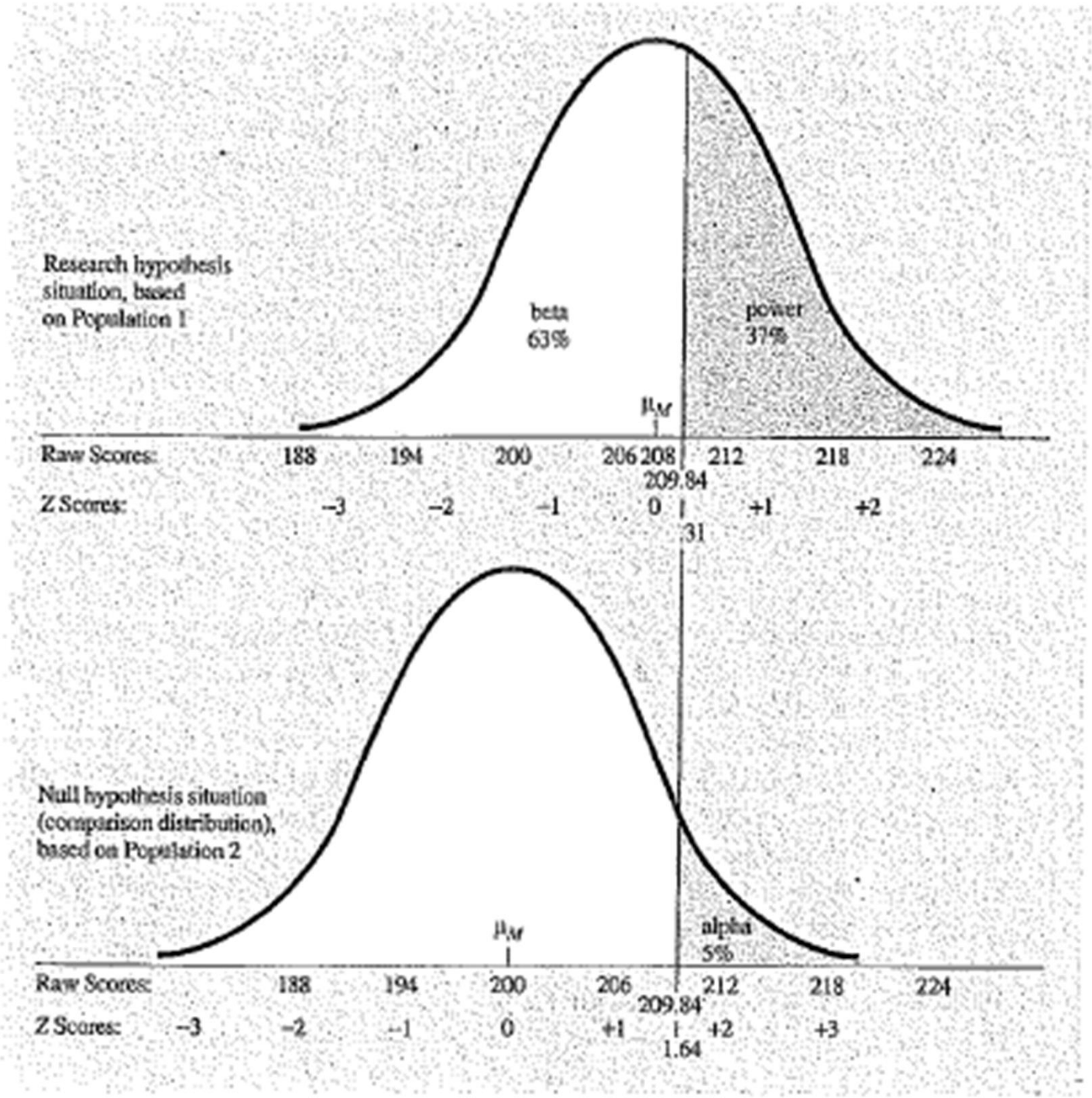
Two group, between subjects, normal populations, standard normal distributions



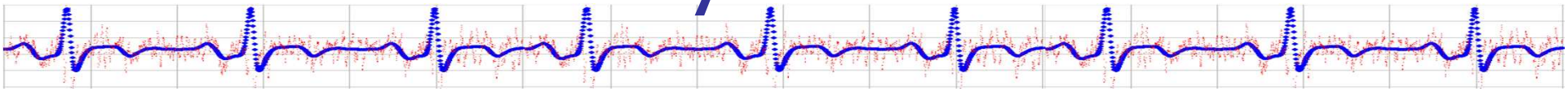
Comparison Population

α (.05)

Z=1.64

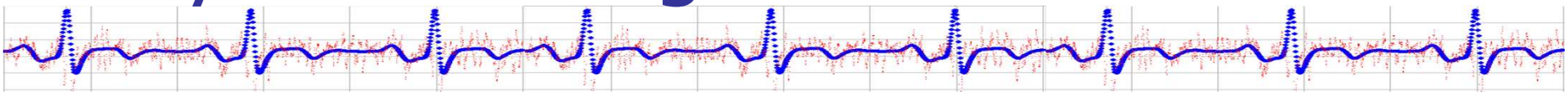

Research hypothesis situation, based on Population 1
beta 63%
power 37%
μ_M
Raw Scores: 188 194 200 206 208 212 218 224
209.84
Z Scores: −3 −2 −1 0 +1 +2
.31
Null hypothesis situation (comparison distribution), based on Population 2
μ_M
alpha 5%
Raw Scores: 188 194 200 206 212 218 224
209.84
Z Scores: −3 −2 −1 0 +1 +2 +3
1.64

Power Analysis



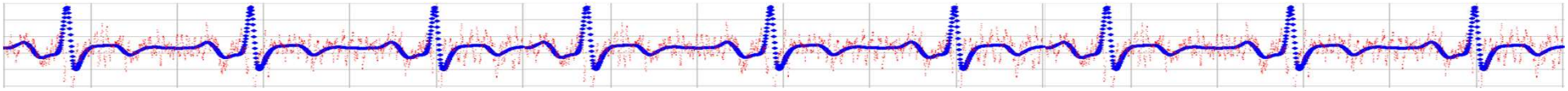
- Should determine number of subjects you need ahead of time by doing a 'power analysis'
- Standard procedure (part of your study plan):
  - Determine statistic you will use
  - Fix alpha and beta (1-power) (and number of tails if appropriate)
  - Estimate expected effect size from prior studies
  - Then: Determine number of subjects you need
- Note: Power
  - Increases with effect size
  - Increases with sample size
  - Decreases with decreasing alpha

Power analyses are different depending on the statistical test you are using…



First up: t-test for independent means.

Effect Size



$$d = \frac{(\mu_1 - \mu_2)}{\sigma}$$

Parameters for population of individuals.
(so, use SD-pooled for t-test of indep means)

Cohen:
d~0.2 small
d~0.5 medium
d~0.8 large

Power table

TABLE 8–4 Approximate Power for Studies Using the *t* Test for Independent Means Testing Hypotheses at the .05 Significance Level

Number of Participants in Each Group	Effect Size		
	Small (.20)	Medium (.50)	Large (.80)
One-tailed test			
10	.11	.29	.53
20	.15	.46	.80
30	.19	.61	.92
40	.22	.72	.97
50	.26	.80	.99
100	.41	.97	*
Two-tailed test			
10	.07	.18	.39
20	.09	.33	.69
30	.12	.47	.86
40	.14	.60	.94
50	.17	.70	.98
100	.29	.94	*

More Useful and Concise
(for practical purposes use a power calculator)

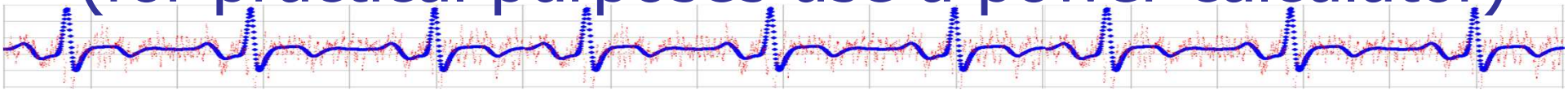
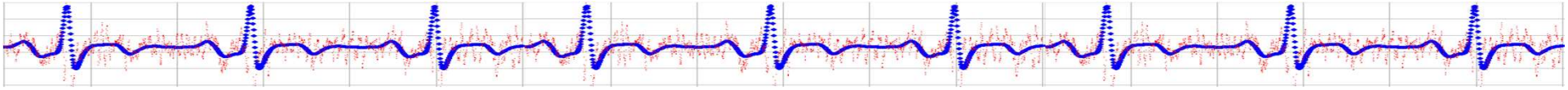


TABLE 8–5 Approximate Number of Participants Needed in Each Group (Assuming Equal Sample Sizes) for 80% Power for the *t* Test for Independent Means, Testing Hypotheses at the .05 Significance Level

	Effect Size		
	Small (.20)	Medium (.50)	Large (.80)
One-tailed	310	50	20
Two-tailed	393	64	26

Power Analysis Exercise



- Based on related research, we expect that there will be a medium effect size in our study of an LED sign in the Food Court affecting wait times.
- How many subjects do we need for a two-group, two-tailed test at α=0.05, 80% power?

More Useful and Concise
(for practical purposes use a power calculator)

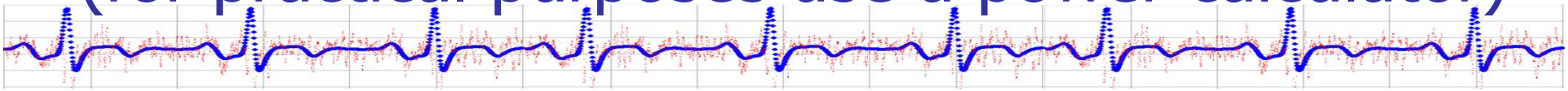
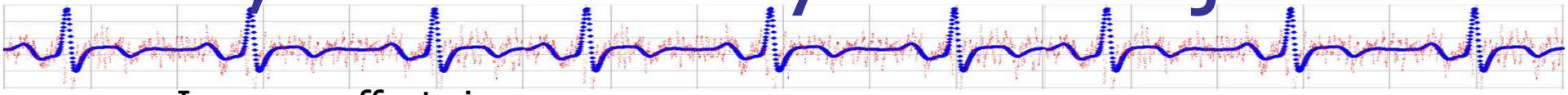


TABLE 8–5 Approximate Number of Participants Needed in Each Group (Assuming Equal Sample Sizes) for 80% Power for the *t* Test for Independent Means, Testing Hypotheses at the .05 Significance Level

	Effect Size		
	Small (.20)	Medium (.50)	Large (.80)
One-tailed	310	50	20
Two-tailed	393	64	26

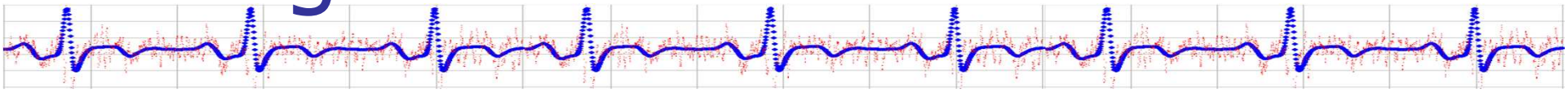
But, I can't study 786 subjects!



- Increase effect size
  - Increase difference in population means (change manipulation)
  - Decrease population variance (better measures, control more extraneous vars)
  - Redesign study to collect many trials of measures per subject
- Relax criteria for Type I error
  - Increase α threshold
  - Change from Two-tailed => one-tailed test
  - *Decreases credibility of your findings*
- Decrease power
  - *Decreases likelihood of getting a significant result*
- Use a different statistic
  - *If possible, maybe consult a statistician*

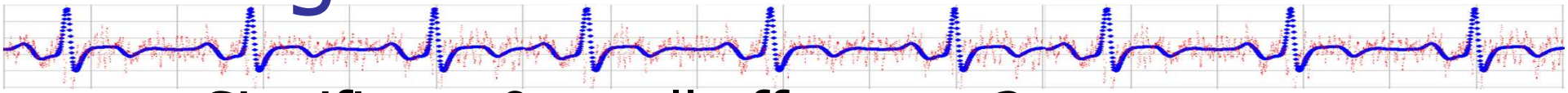
- Practically
  - Usually, redesign experiment so that we have increased effect size or better measures for decreased variance
  - OR, call it a "pilot study"

Interpreting results: Significance & Effect Size



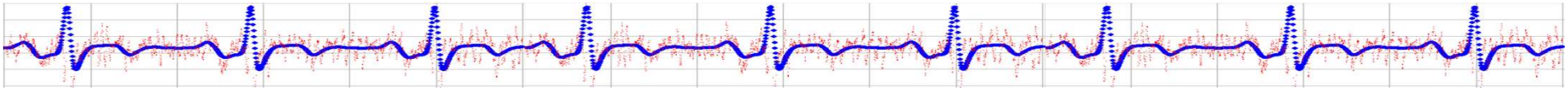
- Significance
  - Just indicates that it is likely there is a non-zero difference between populations.
  - Says nothing about how big the difference is.
- Effect Size
  - Only meaningful if result is significant.
  - Indicates how big the difference is (usually normalized to number of std-deviations)

Interpreting results: Significance & Effect Size



- Significant & small effect => ?
  - Real difference, but slight.
  - Probably not of practical importance.
- Significant & large effect => ?
  - Real difference, likely meaningful.
- Significant & small sample => ?
  - Significant & possibly important.
- Non-significant & small sample => ?
  - Inconclusive
- Non-significant & large sample => ?
  - Evidence there really is no difference

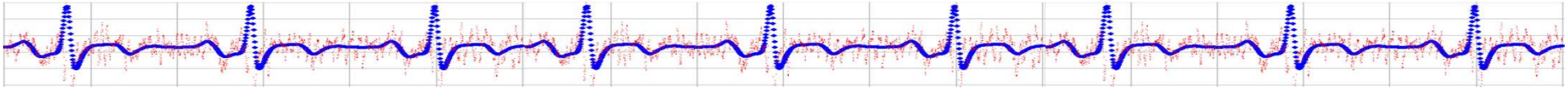
Group Exercise



- Compute effect size for the study
- Characterize as small/medium/large

- You are now going to do a follow-up study using similar interventions and measures (ie assume same effect size)
- Do a power analysis to determine how many subjects you would need for a two-group between-subjects experiment with 80% power, alpha=0.05, two-tailed test.

Power & Effect Size for Correlation



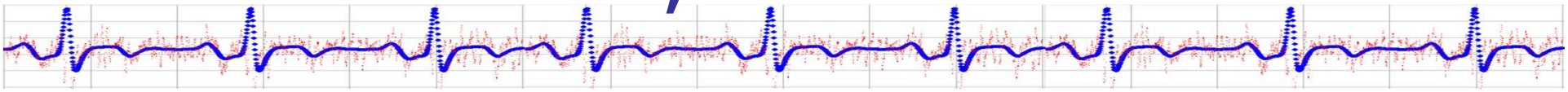
- Effect size = |r|
- Power, see table 11-7, pg 465 Aron
  - Usually, given
    - Expected effect size
    - Test criteria
      - Desired significance level (usually 0.05)
      - Desired power (usually 0.8)
      - Directionality of test

Table 11-7 Approximate Power of Studies Using the Correlation Coefficient (r) for Testing Hypotheses at the .05 Level of Significance

		Effect Size		
		Small ($r = .10$)	Medium ($r = .30$)	Large ($r = .50$)
Two-tailed				
Total N:	10	.06	.13	.33
	20	.07	.25	.64
	30	.08	.37	.83
	40	.09	.48	.92
	50	.11	.57	.97
	100	.17	.86	*
One-tailed				
Total N:	10	.08	.22	.46
	20	.11	.37	.75
	30	.13	.50	.90
	40	.15	.60	.96
	50	.17	.69	.98
	100	.26	.92	*

*Power is nearly 1.

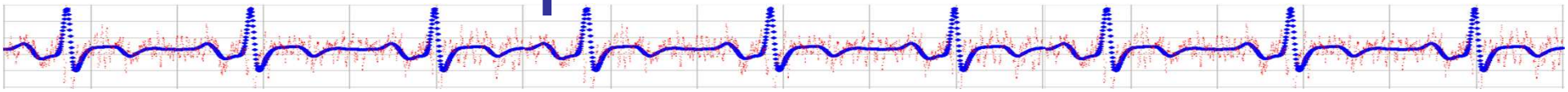
Table 11-8, Aron



Approximate number of participants needed for 80% power for a study using the correlation coefficient (r) for testing a hypothesis at the .05 significance level

Effect size		
Small (r=0.1)	**Medium (r=0.3)**	**Large (r=0.5)**
783	**85**	**28**

Effect Size & Power for X^2 test for independence



- Completely different formulas than for Pearson r or t-test.
- Dependent on df.
- For 2x2, effect size = "phi"

$$\sqrt{\frac{x^2}{N}}$$

Effect Size & Power for X^2

Table 13-10 Approximate Total Number of Participants Needed for 80% Power for the Chi-Square Test for Independence for Testing Hypotheses at the .05 Significance Level

	Effect Size		
Total *df*	Small	Medium	Large
1	785	87	26
2	964	107	39
3	1,090	121	44
4	1,194	133	48

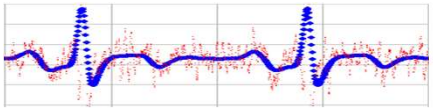
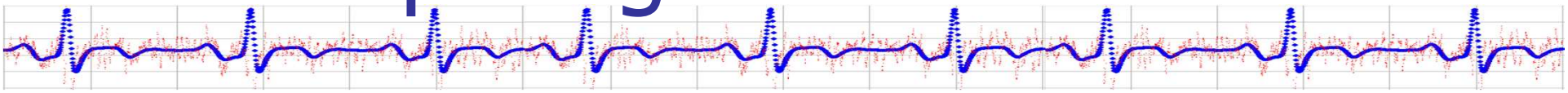


Table 13-9 Approximate Power for the Chi-Square Test for Independence for Testing Hypotheses at the .05 Significance Level

Total *df*	Total *N*	Effect Size		
		Small	Medium	Large
1	25	.08	.32	.70
	50	.11	.56	.94
	100	.17	.85	*
	200	.29	.99	*
2	25	.07	.25	.60
	50	.09	.46	.90
	100	.13	.77	*
	200	.23	.97	*
3	25	.07	.21	.54
	50	.08	.40	.86
	100	.12	.71	.99
	200	.19	.96	*
4	25	.06	.19	.50
	50	.08	.36	.82
	100	.11	.66	.99
	200	.17	.94	*

*Nearly 1.

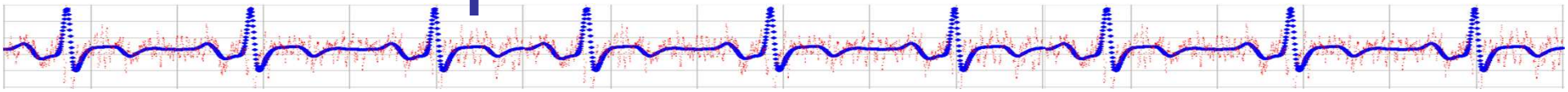
Computing effect size



- Some thoughtless authors do not include means & stddevs (per group) in their article…
- R package 'compute.es' contains a variety of methods for computing effect size given other info (e.g., t score, N1, N2)

- Morale: Always include means & stddevs
- Better: Report effect sizes yourself!

Chi-Squared

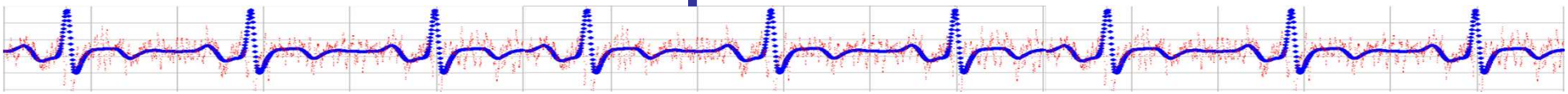


- https://www.anzmtg.org/stats/PowerCalculator/PowerChiSquare

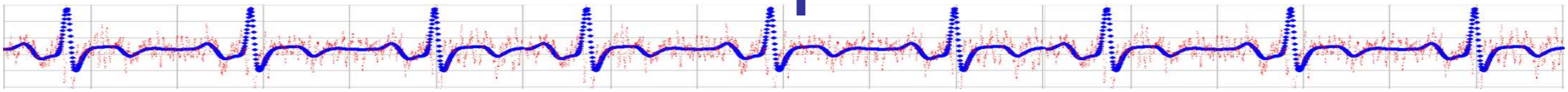
G*Power


Not secure | www.gpower.hhu.de
HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF
Allgemeine Psychologie und Arbeitspsychologie
Portale
ULB-Katalog
Intranet
HHU Start › Fakultäten › MNF › Fächer › Psychologie › Arbeitsgruppen › Allgemeine Psychologie und Arbeitspsychologie › G*Power
Zurück zur Psychologie
Startseite
Arbeitsgruppe
Themen für Abschlussarbeiten
Über Abschluss- und Hausarbeiten
Über Referate
Unsere Forschung
Publikationen
G*Power
ART
APriot
G*Power: Statistical Power Analyses for Windows and Mac
G*Power is a tool to compute statistical power analyses for many different t tests, F tests, χ2 tests, z tests and some exact tests. G*Power can also be used to compute effect sizes and to display graphically the results of power analyses.
HHU Suche
Screenshots (click to enlarge)
Main Window
Main Window (Table)
Power Plot
Power Plot (Table)
Register
Whenever we find a problem with G*Power we provide an update as quickly as we can. We will inform you about updates if you click here and add your e-mail address to our mailing list. We will only use your e-mail address to inform you about updates. We will not use your e-mail address for other purposes. We will not give your e-mail address to anyone else. You can withdraw your e-mail address from the mailing list at any time.
Help
Download the G*Power manual (PDF)

t-test concepts

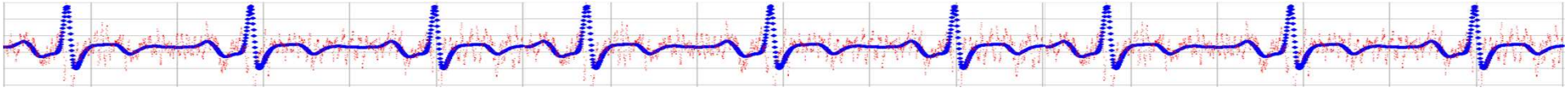


t-test for independent means



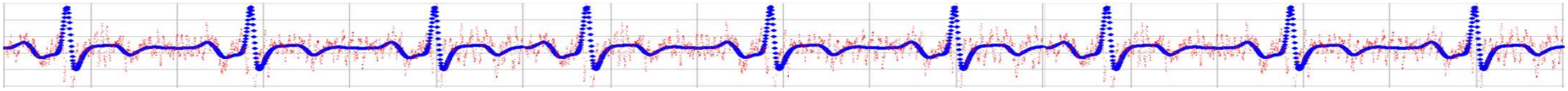
- Tests association between binomial IV and numeric DV.
- Examples:
  - WizziWord vs. Word => wpm
  - Small vs. Large Monitors => wpd
  - Wait time sign vs. none => satisfaction

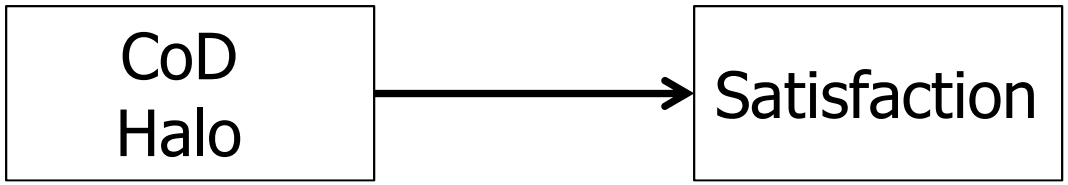
Understanding numeric measures



- Sources of variance
  - IV
  - Other uncontrolled factors (“error variance”)

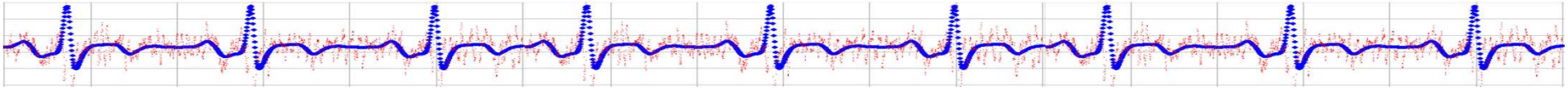
Example: Call of Duty vs. Halo





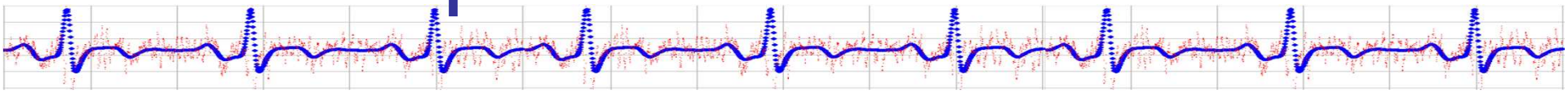
- What variables might affect Satisfaction?
- Typically, one subject's Satisfaction score = **TrueSatisfaction** + var1 + var2 + var3 + …
- A sum of random variables.

Central Limit Theorem



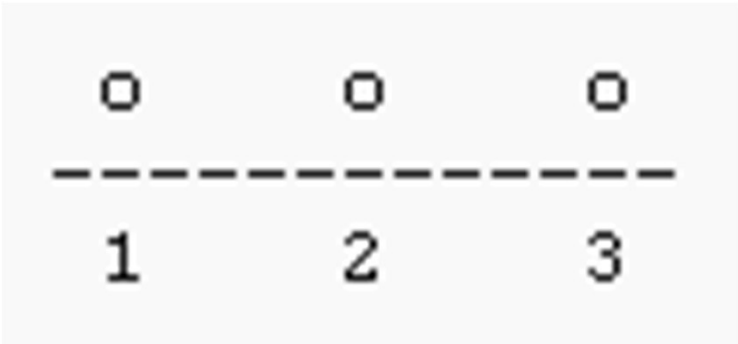
- If (many) independent, random variables with the same distribution are added, the result is approximately a normal curve

Why be normal? Example



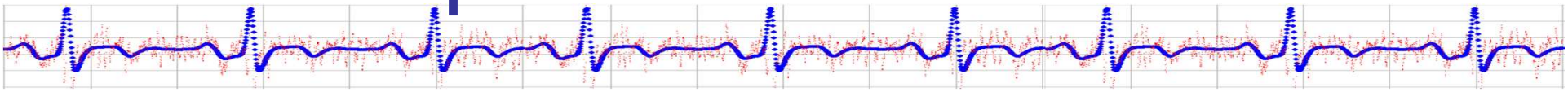
- Suppose random variable X has distribution

$$X = \begin{cases} 1 & \text{with probability } 1/3, \\ 2 & \text{with probability } 1/3, \\ 3 & \text{with probability } 1/3. \end{cases}$$



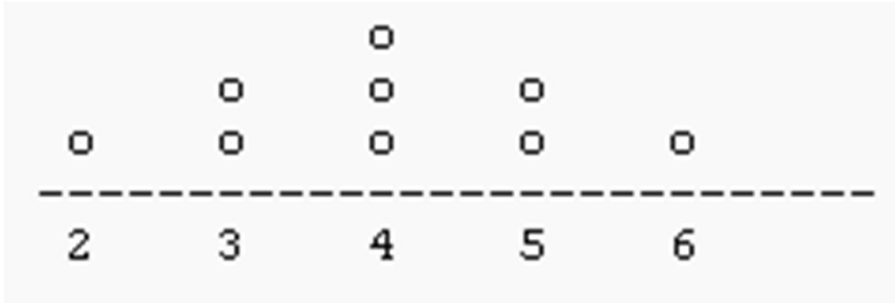
From wikipedia

Why be normal? Example

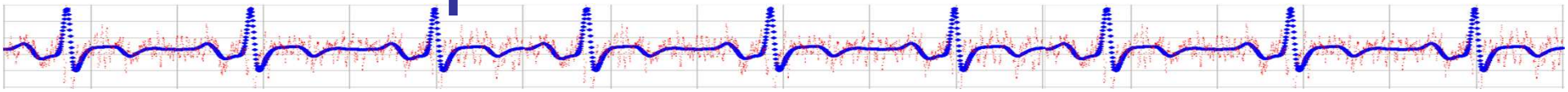


- Now, consider the distribution of X+X

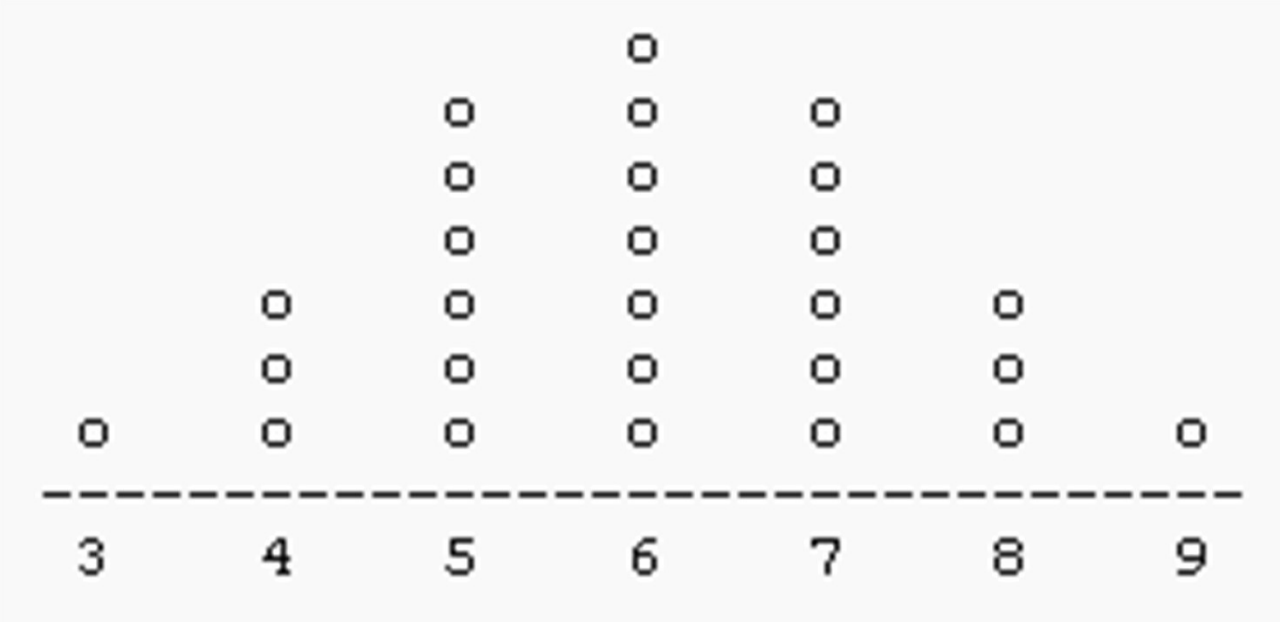
$$\left\{\begin{array}{rcl} 1+1 &=& 2 \\ 1+2 &=& 3 \\ 1+3 &=& 4 \\ 2+1 &=& 3 \\ 2+2 &=& 4 \\ 2+3 &=& 5 \\ 3+1 &=& 4 \\ 3+2 &=& 5 \\ 3+3 &=& 6 \end{array}\right\} = \left\{\begin{array}{ll} 2 & \text{with probability } 1/9 \\ 3 & \text{with probability } 2/9 \\ 4 & \text{with probability } 3/9 \\ 5 & \text{with probability } 2/9 \\ 6 & \text{with probability } 1/9 \end{array}\right\}$$



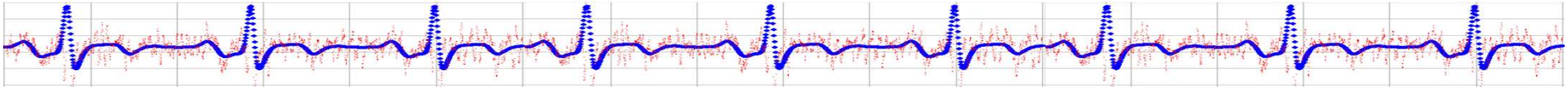
Why be normal? Example



- Now, consider the distribution of X+X+X

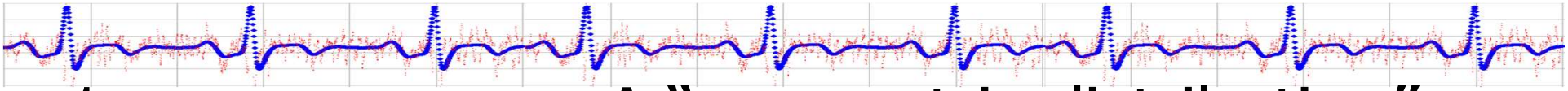


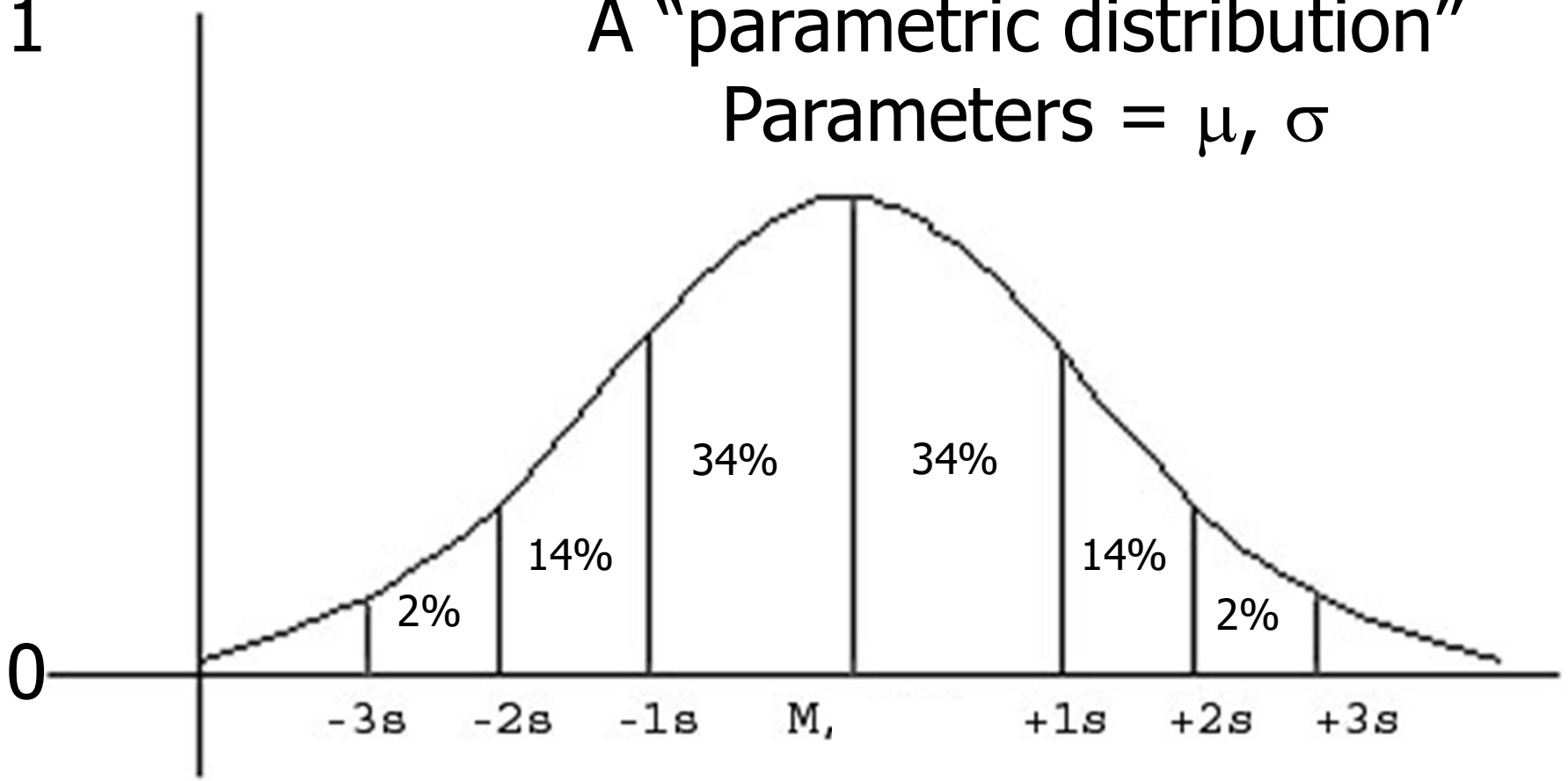
Central Limit Theorem



- The sum of independent random variables with the same underlying distribution will be approximately normal

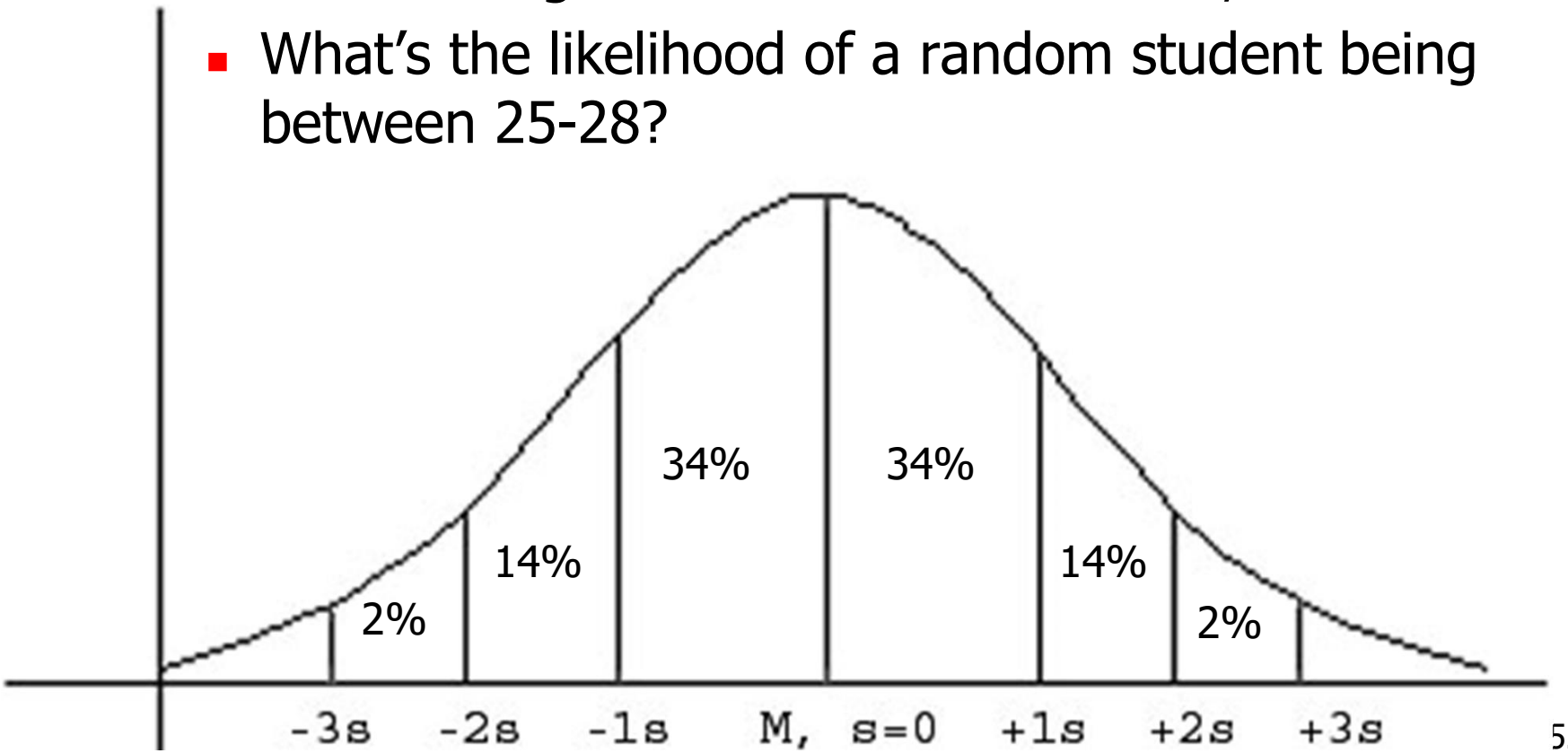
The Normal Curve



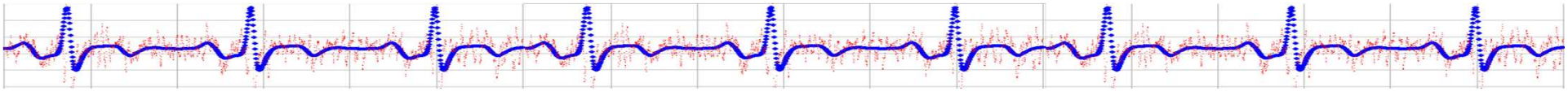


Estimating probabilities using the normal curve

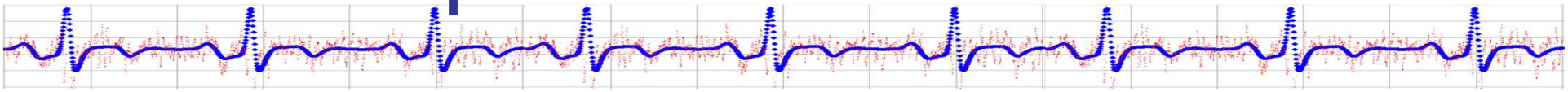
- p(x>sample>y)
  - Ex: know age of NU students is M=25, SD=3
  - What's the likelihood of a random student being between 25-28?



Given what we know so far.. We can already do some hypothesis testing!

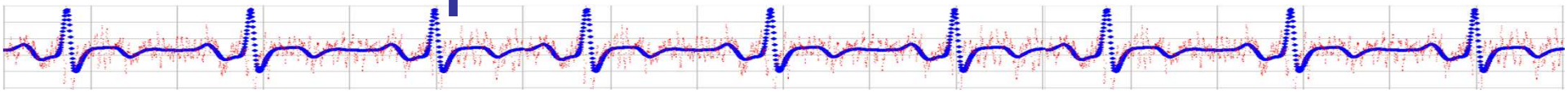


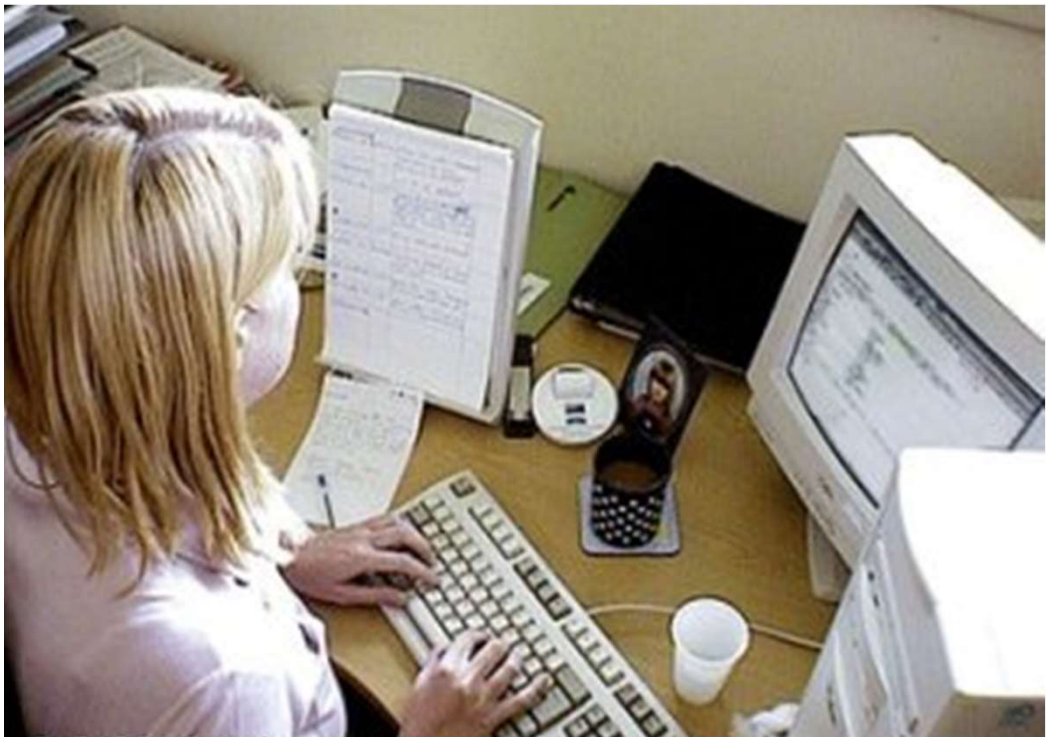
Example

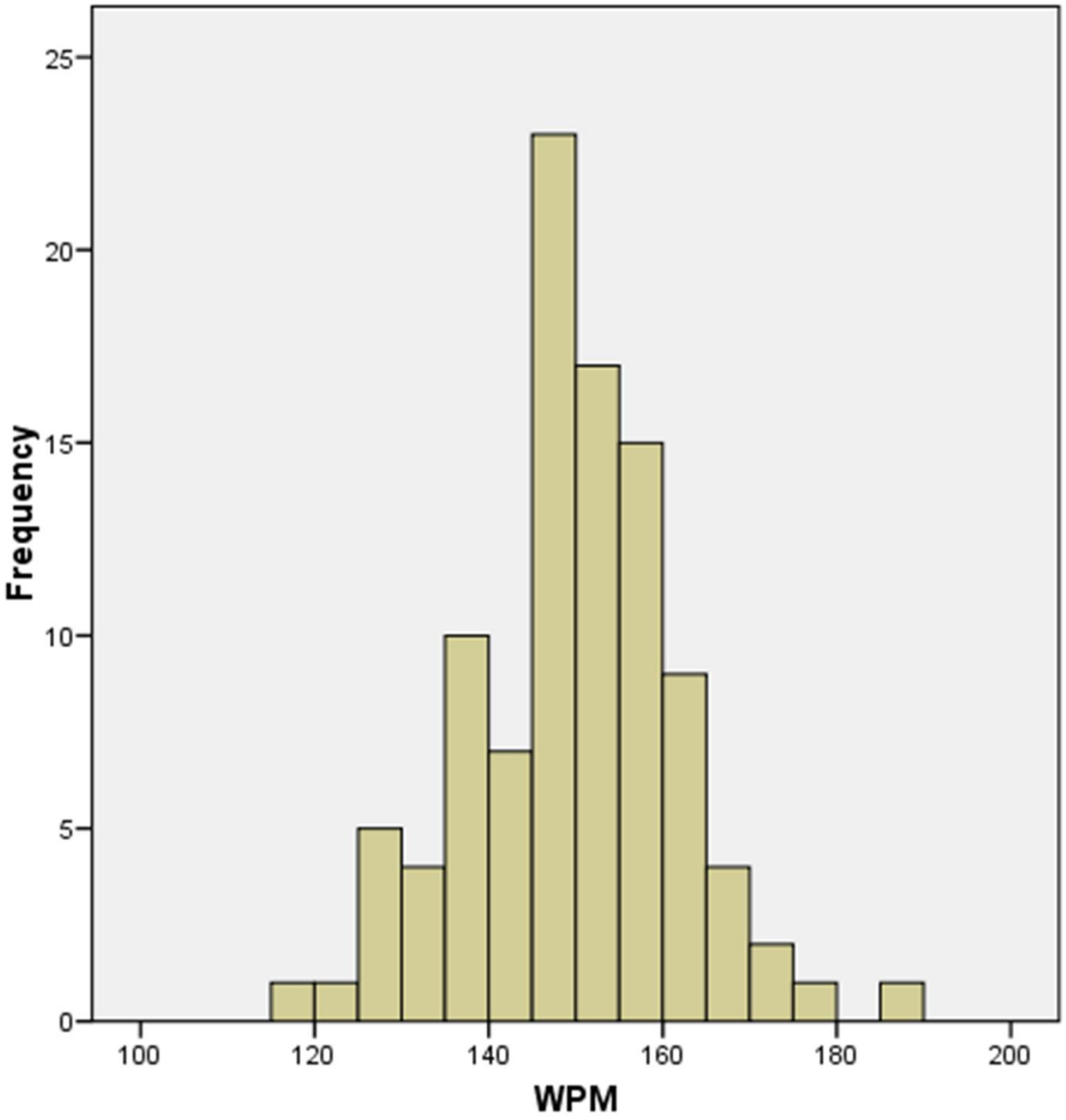


- You have 100 admins in your company.
- They all use Word.
- You want to consider changing to WizziWord.
- Hypothesize it will increase their net productivity, measured as word per minute typed, averaged over an entire day.

Example

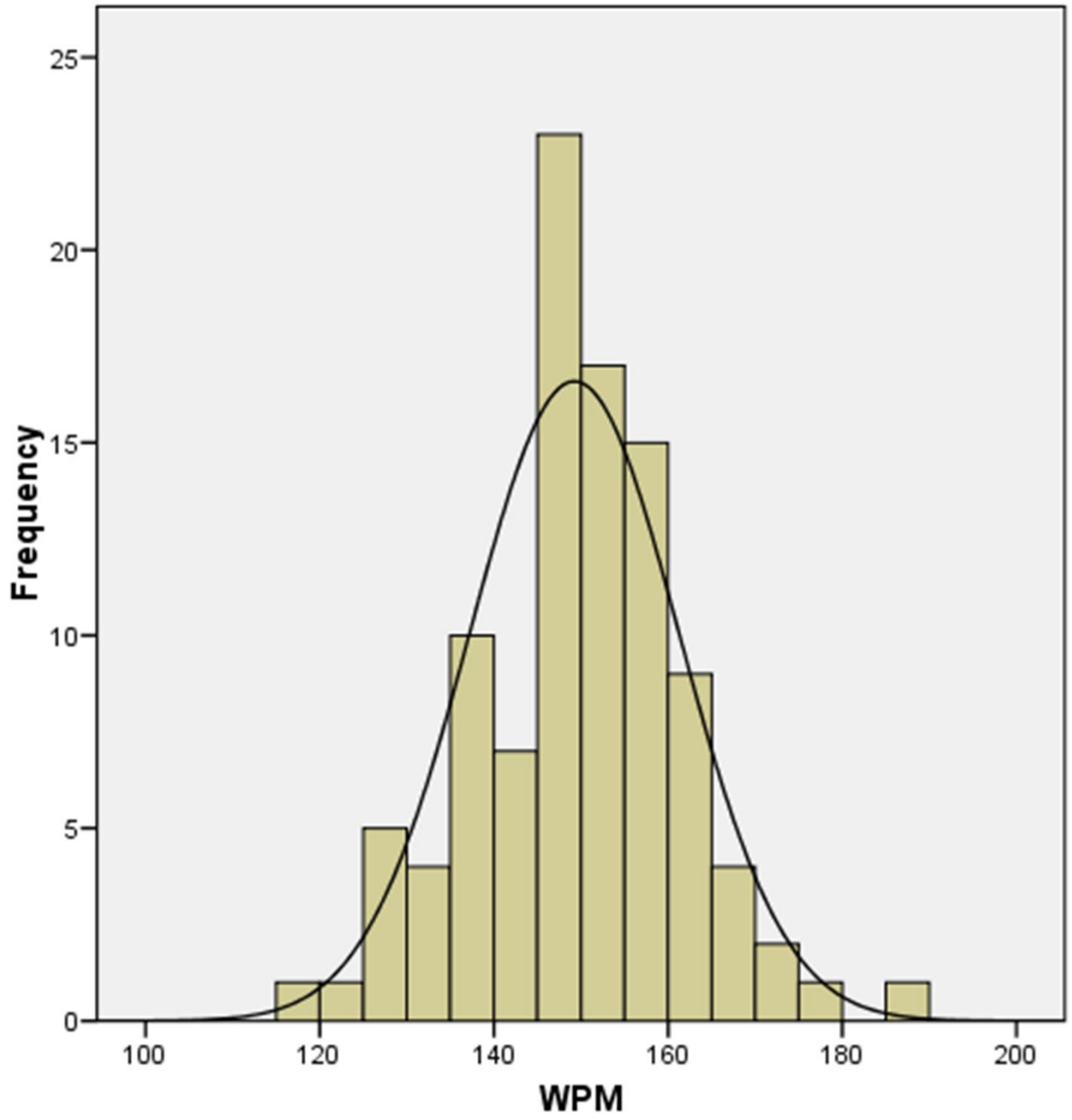






Productivity of
ALL admins
BEFORE WizziWord

Mean =149.23
Std. Dev. =12.02
N =100



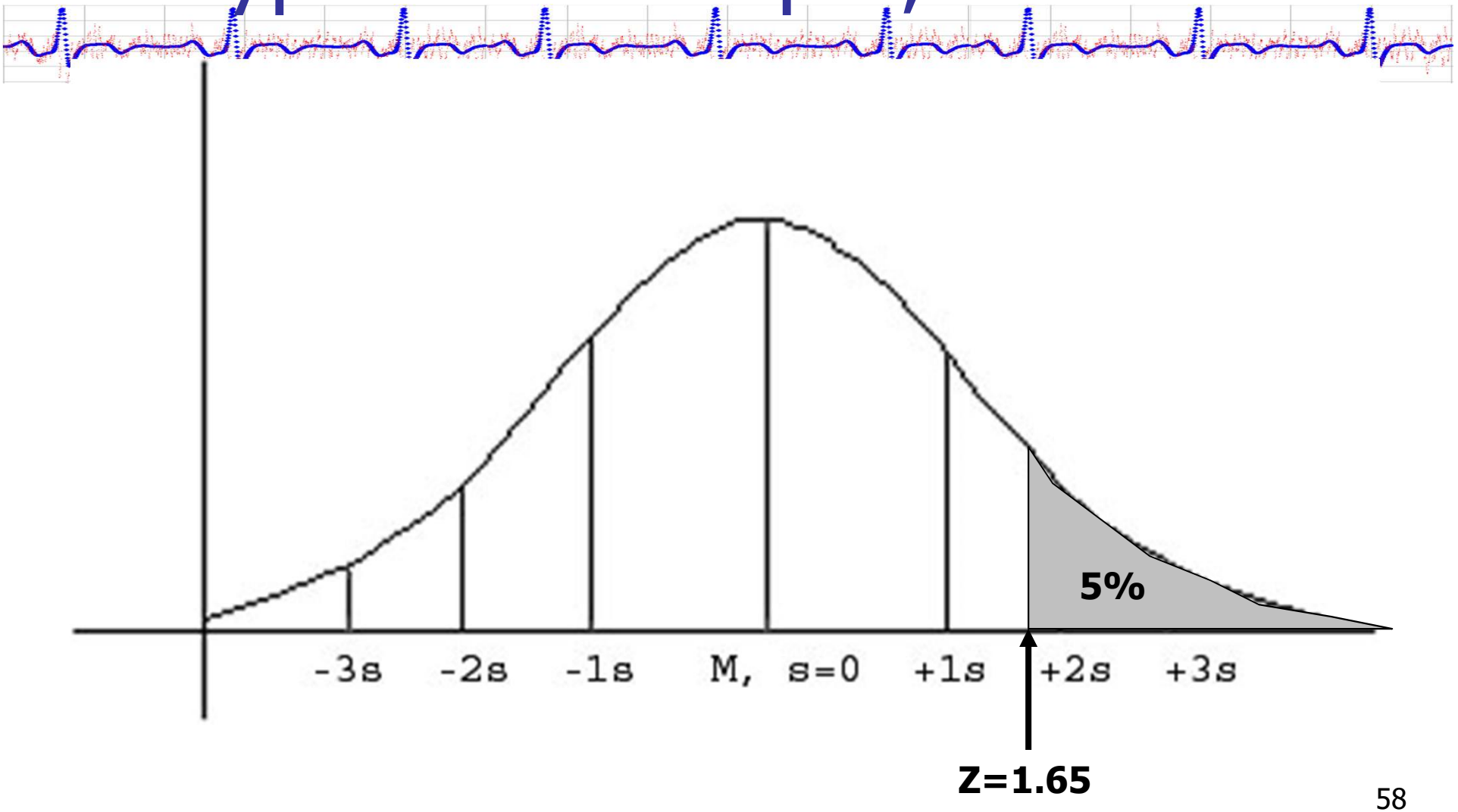
Experiment:
Give 1 admin WW
for a month

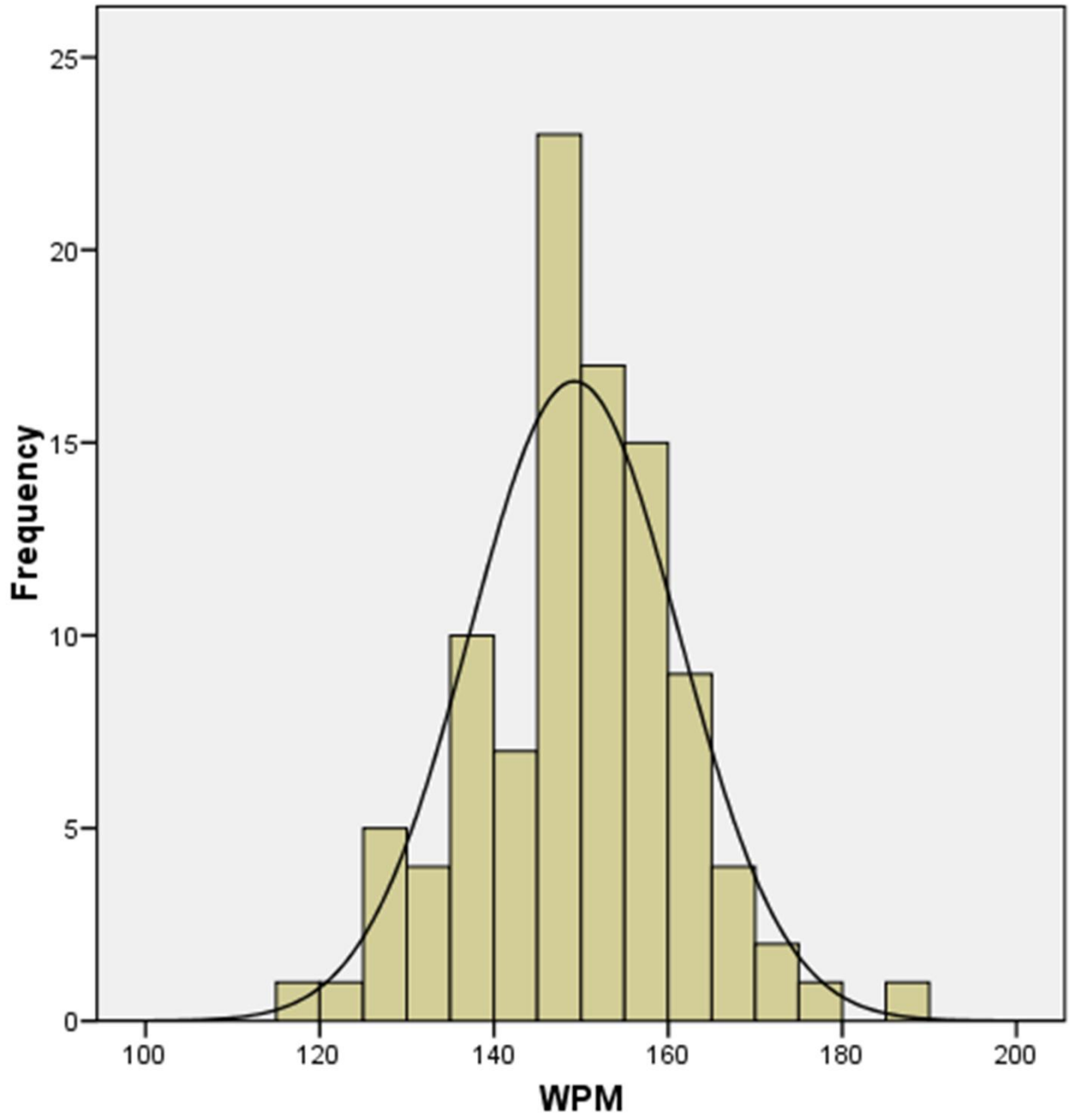
types at 160 wpm

H1: WW > Word

Z=0.9

Example: 1 admin using WW types at 160 wpm, 1 tail

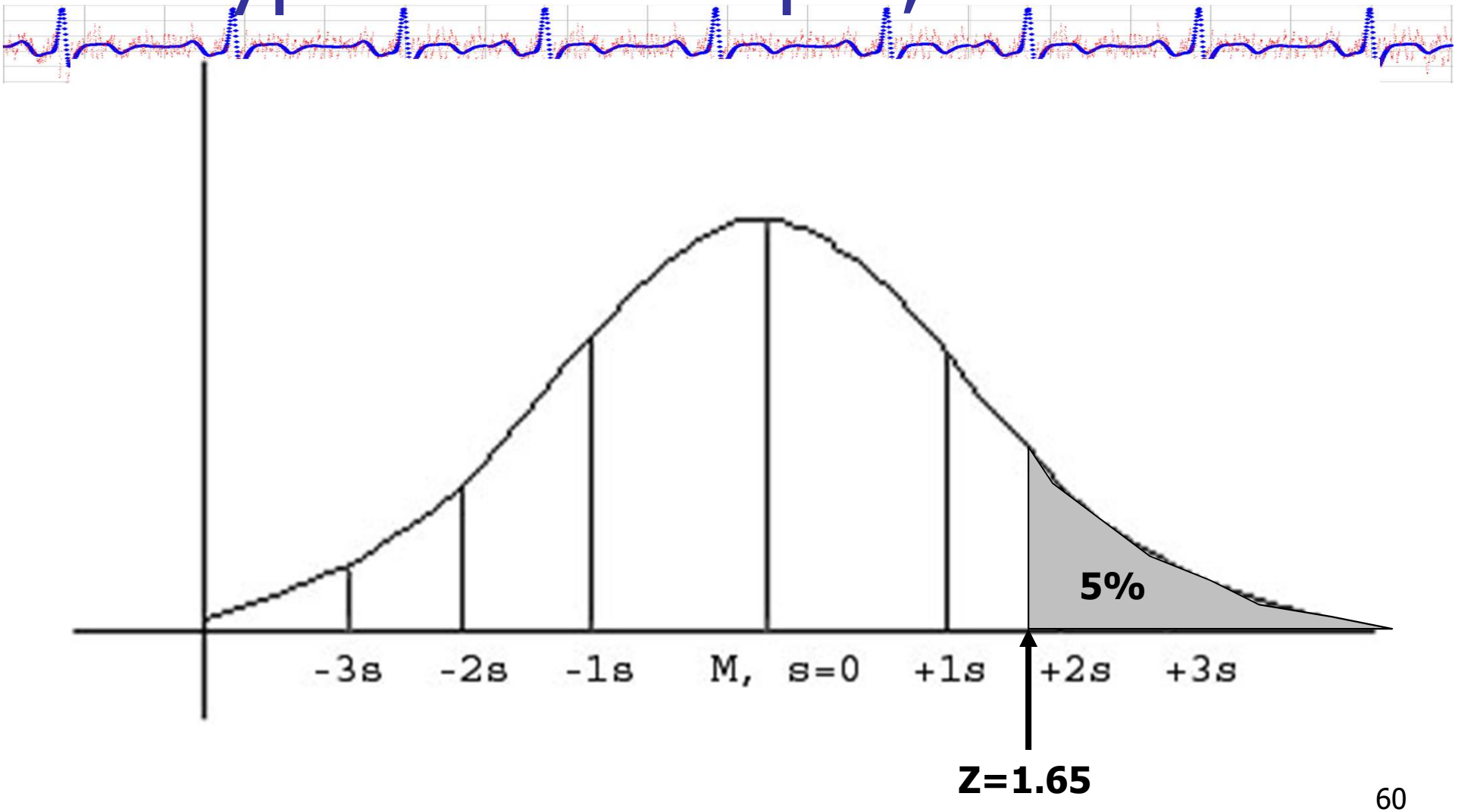


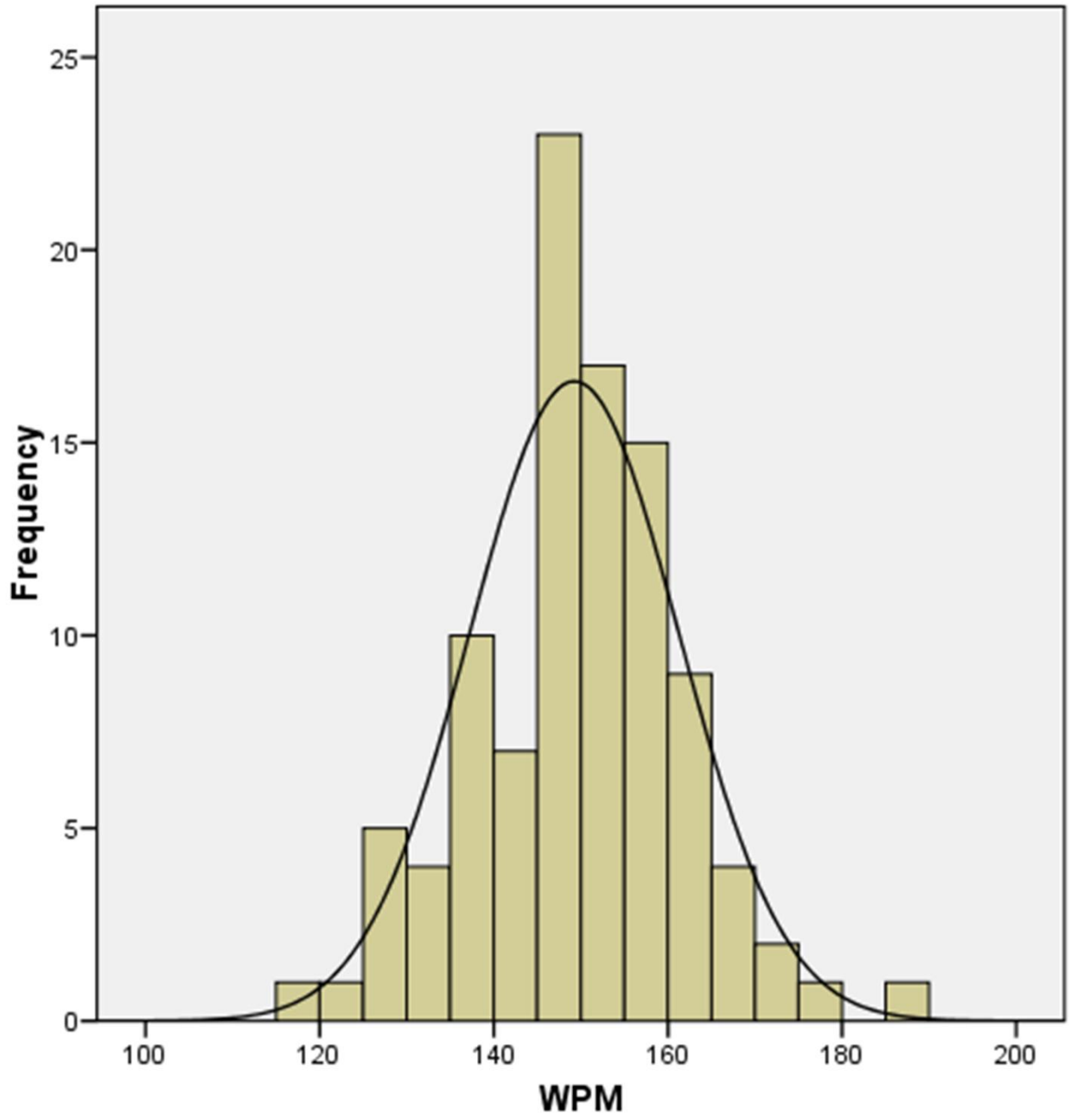


Example:
1 admin using WW
types at 190 wpm
1 tail test

Z=3.4

Example: 1 admin using WW types at 190 wpm, 1 tail

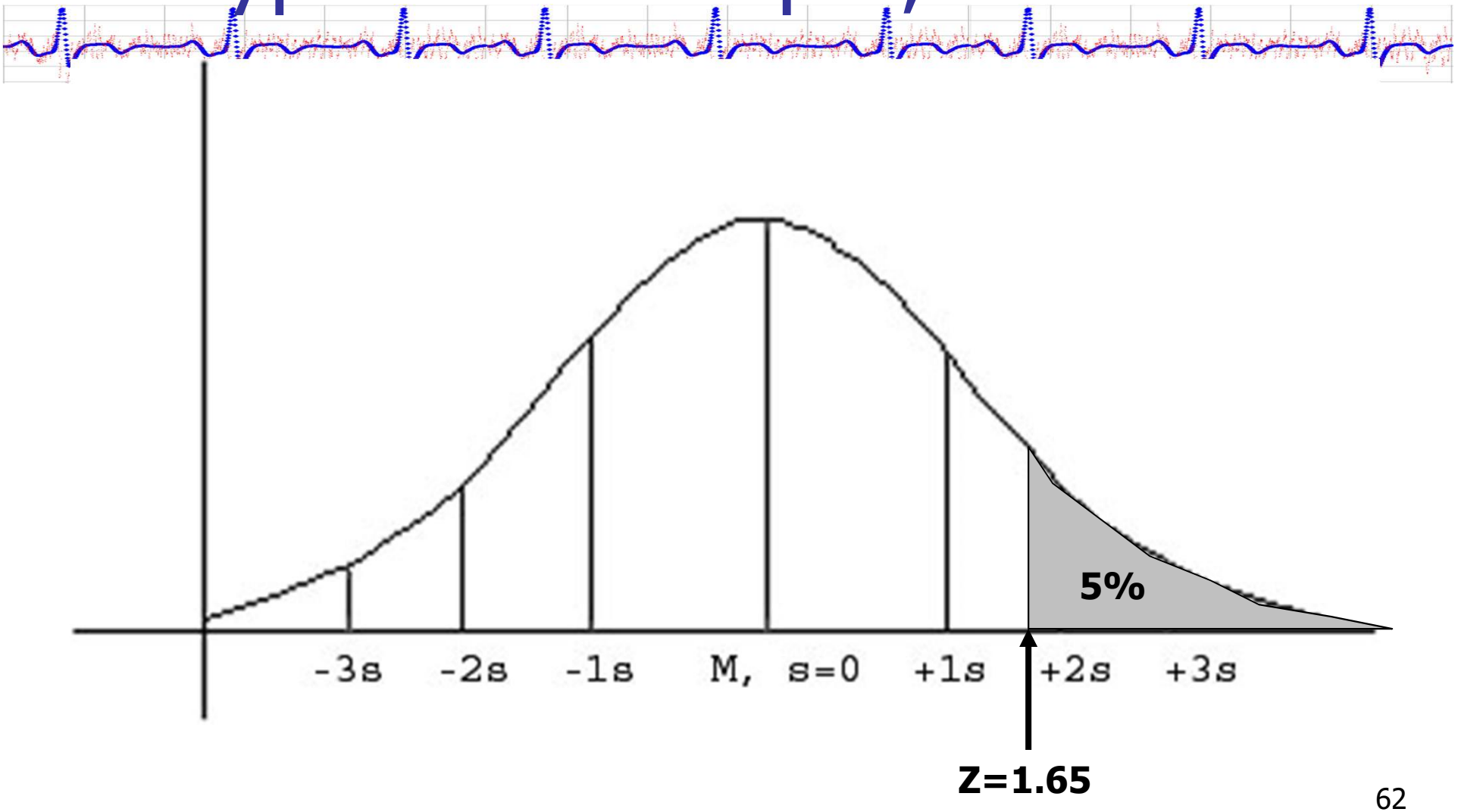


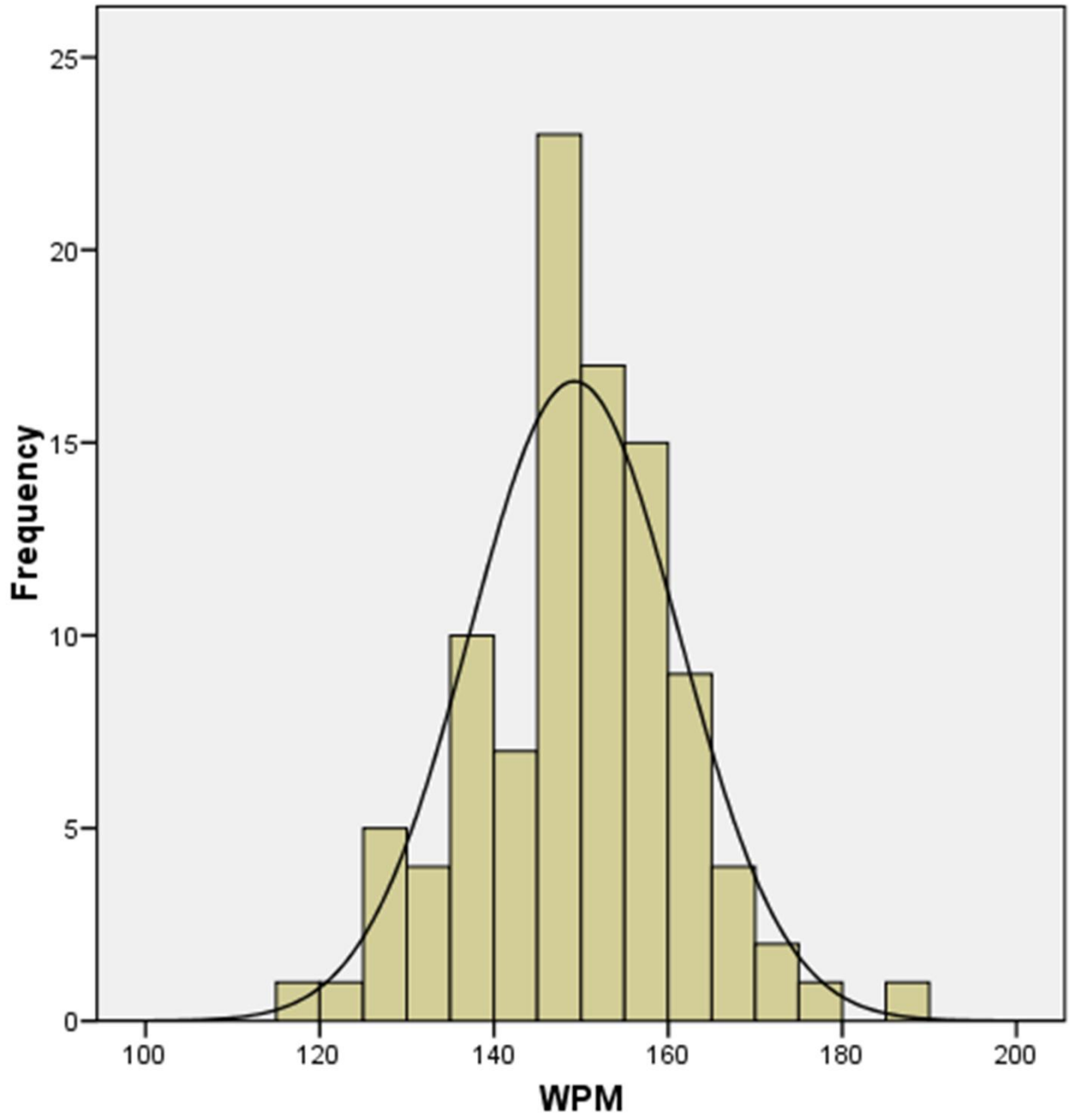


Example:
1 admin using WW
types at 121 wpm
1 tail test

Z=-2.3

Example: 1 admin using WW types at 121 wpm, 1 tail



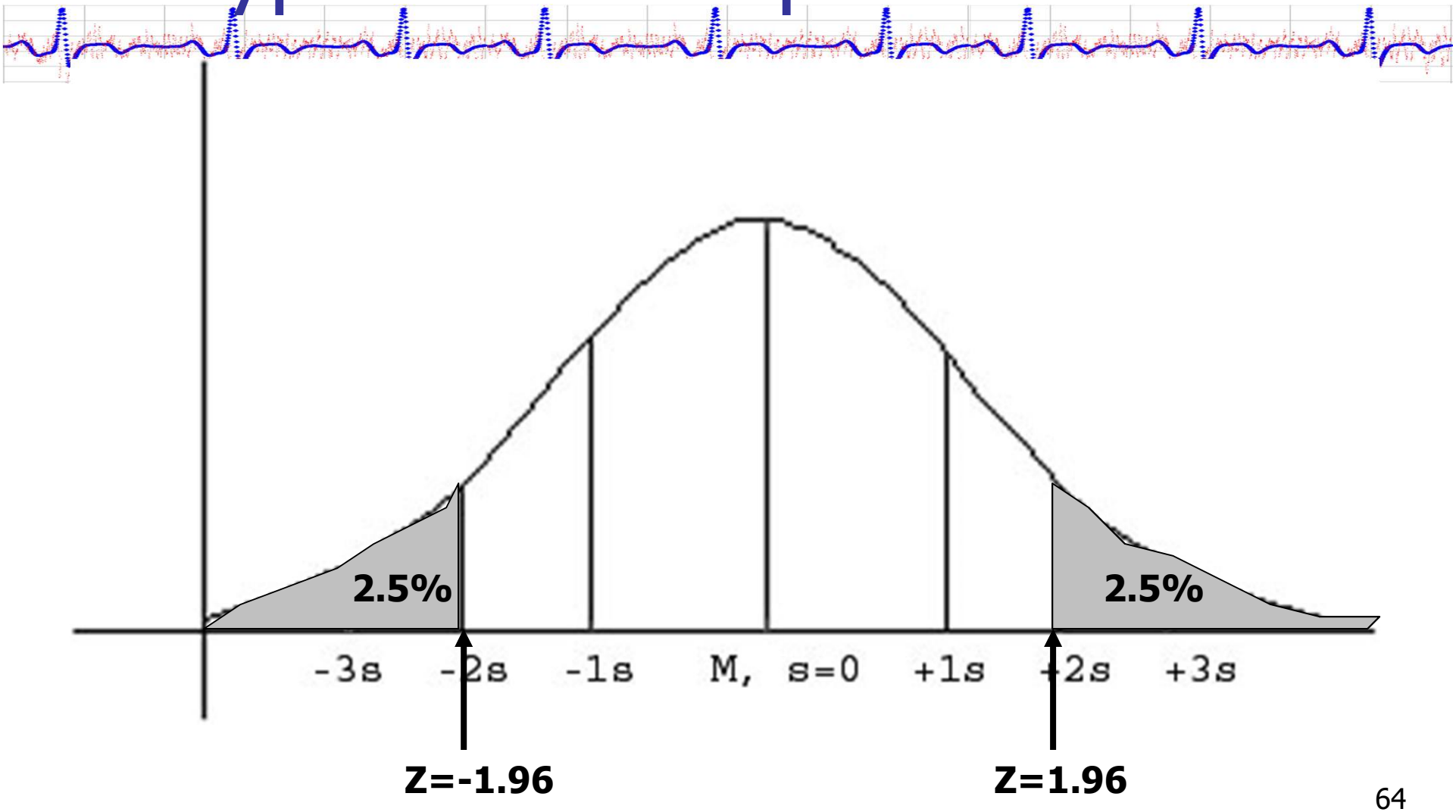


Example:
1 admin using WW types at 121 wpm

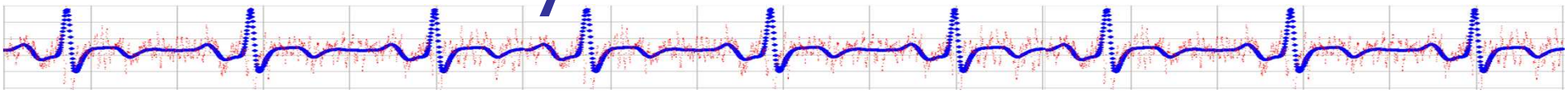
H1: WW ≠ W

Z=-2.3

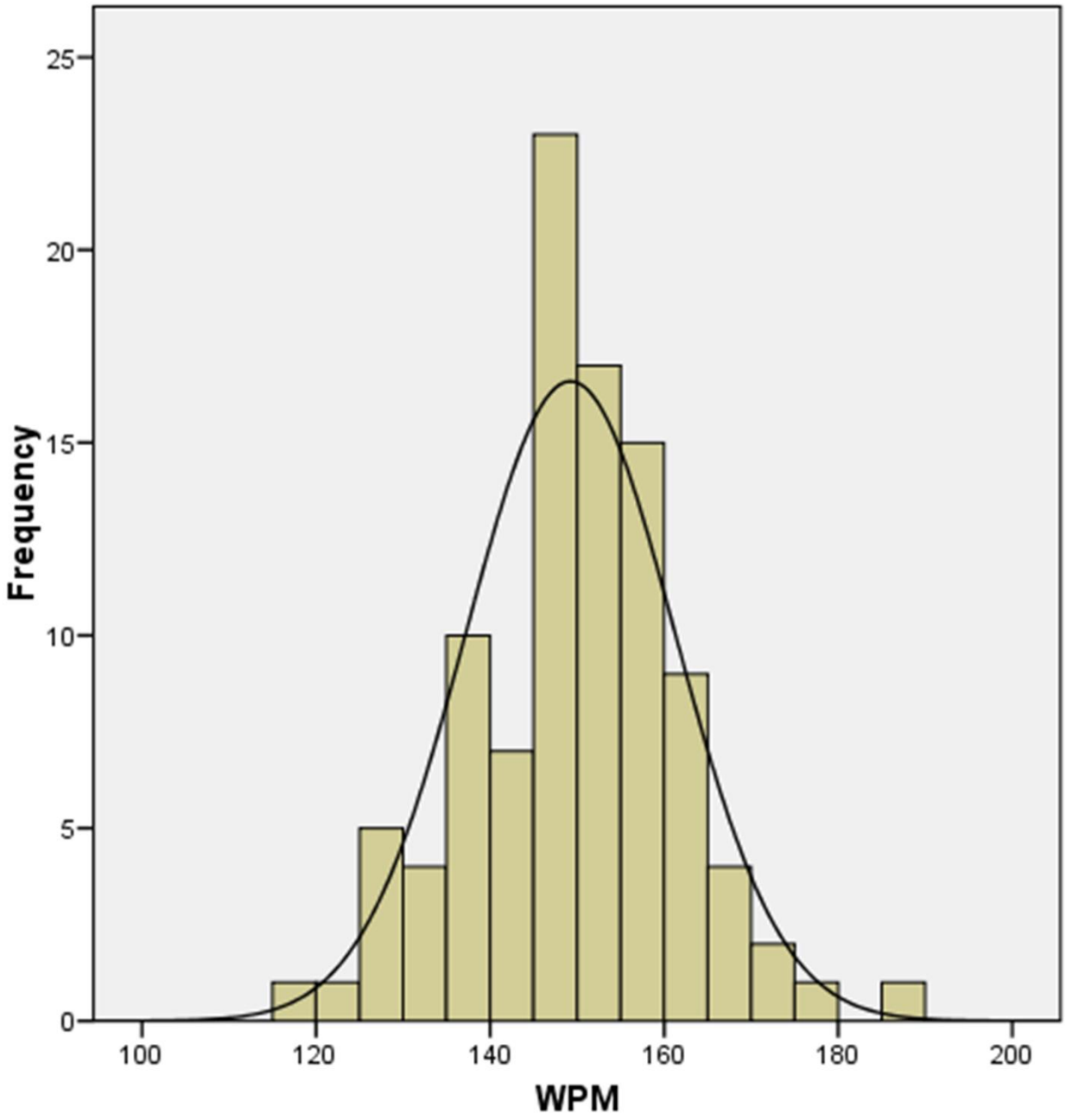
Example: 1 admin using WW types at 121 wpm 2 tail test



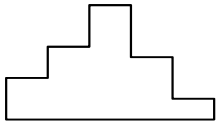
Don't try this at home



- You would never do a study this way.
- Why?
  - Can't control extraneous variables through randomization.
  - Usually don't know population statistics.
  - Can't generalize from an individual.

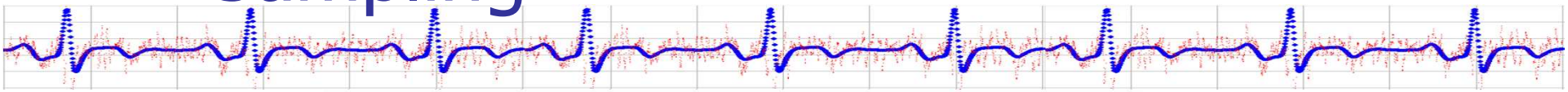


Example:
Take sample of 10 admins, give WW, assess their typing speeds



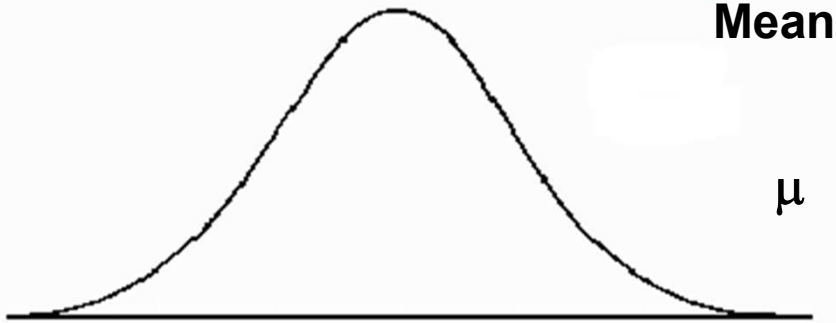
Now what?

Sampling



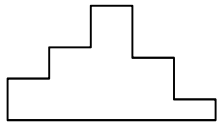
Mean? **Variance?**

Population



μ σ^2

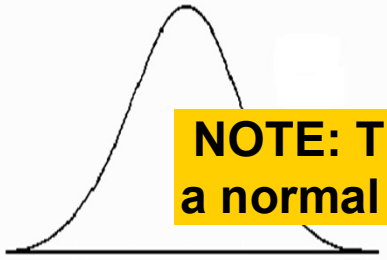
Sample of size N



$$M = \frac{\sum X}{N} \qquad SD^2 = \frac{\sum (X - M)^2}{N - 1}$$

Mean values from all possible samples of size N

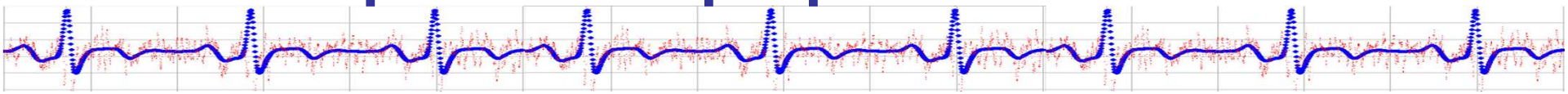
aka "distribution of means"



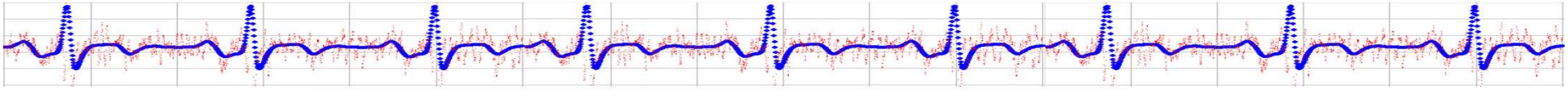
NOTE: This is a normal curve

μ $\sigma_M^2 = \frac{\sigma^2}{N}$

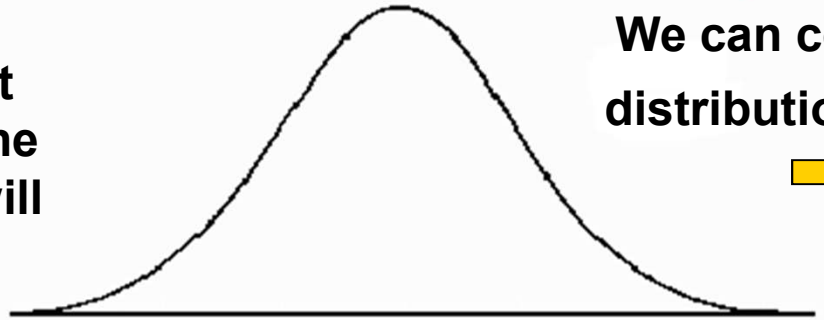
Single sample test, with sample size>1, and known comparison population…



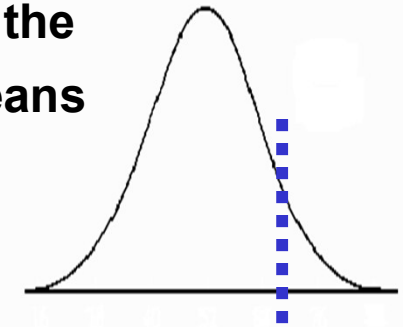
Hypothesis testing with a sample wrt distribution of means



Given info about population and the sample size we will be using (N)

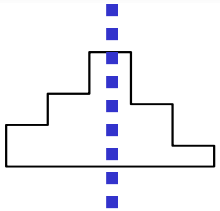


We can compute the distribution of means

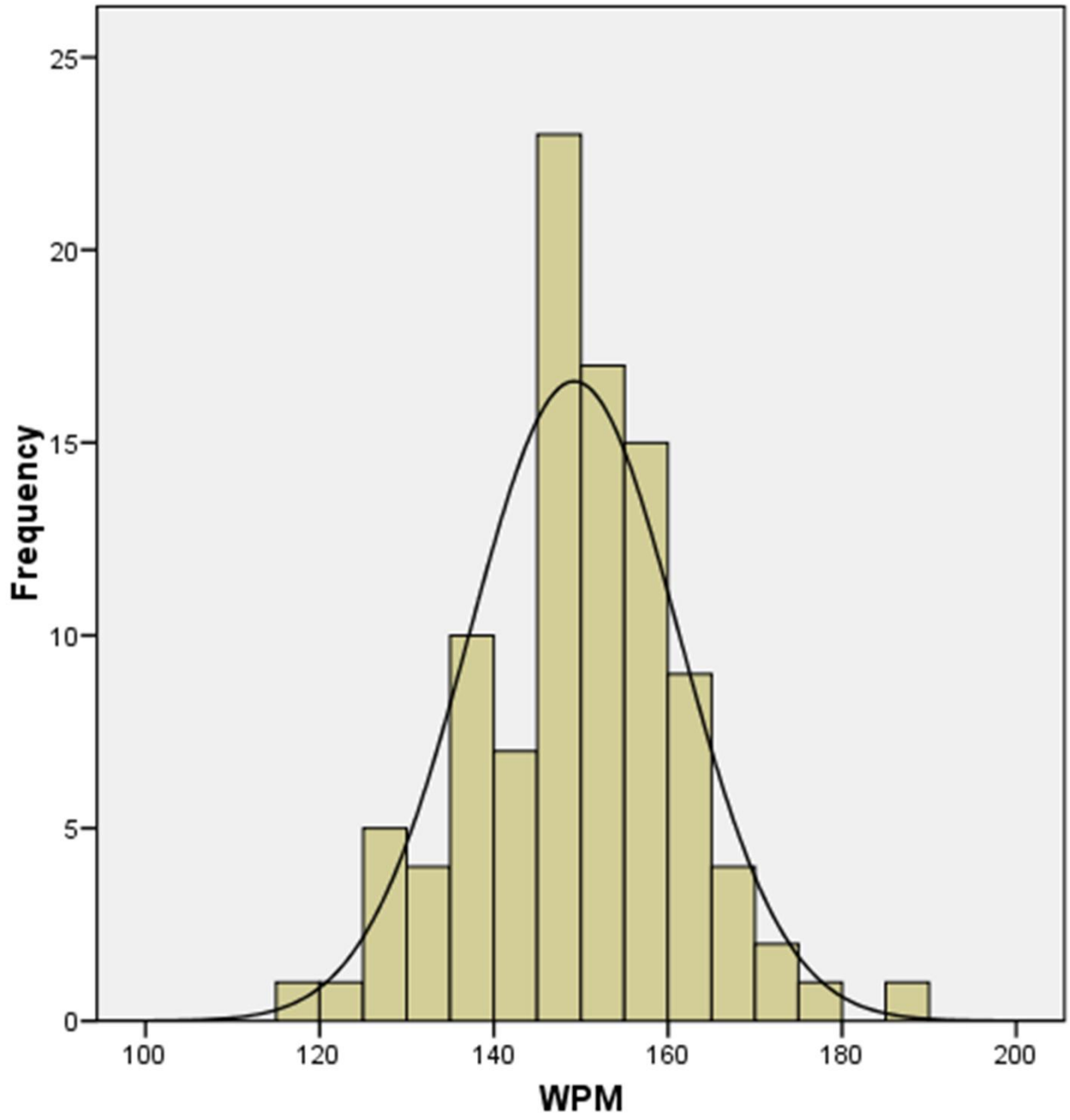


and finally determine the probability that this mean occurred by chance

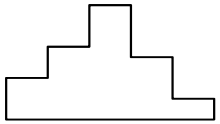
Now, given a particular sample of size N



We compute its mean

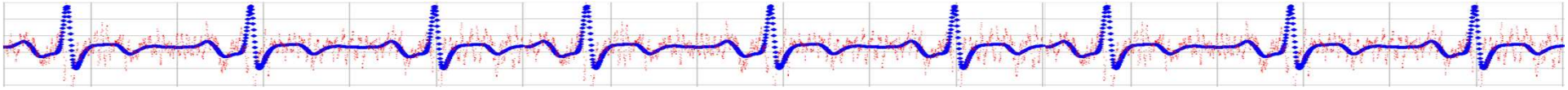


Example:
Take sample of 10 admins, give WW, assess their typing speeds



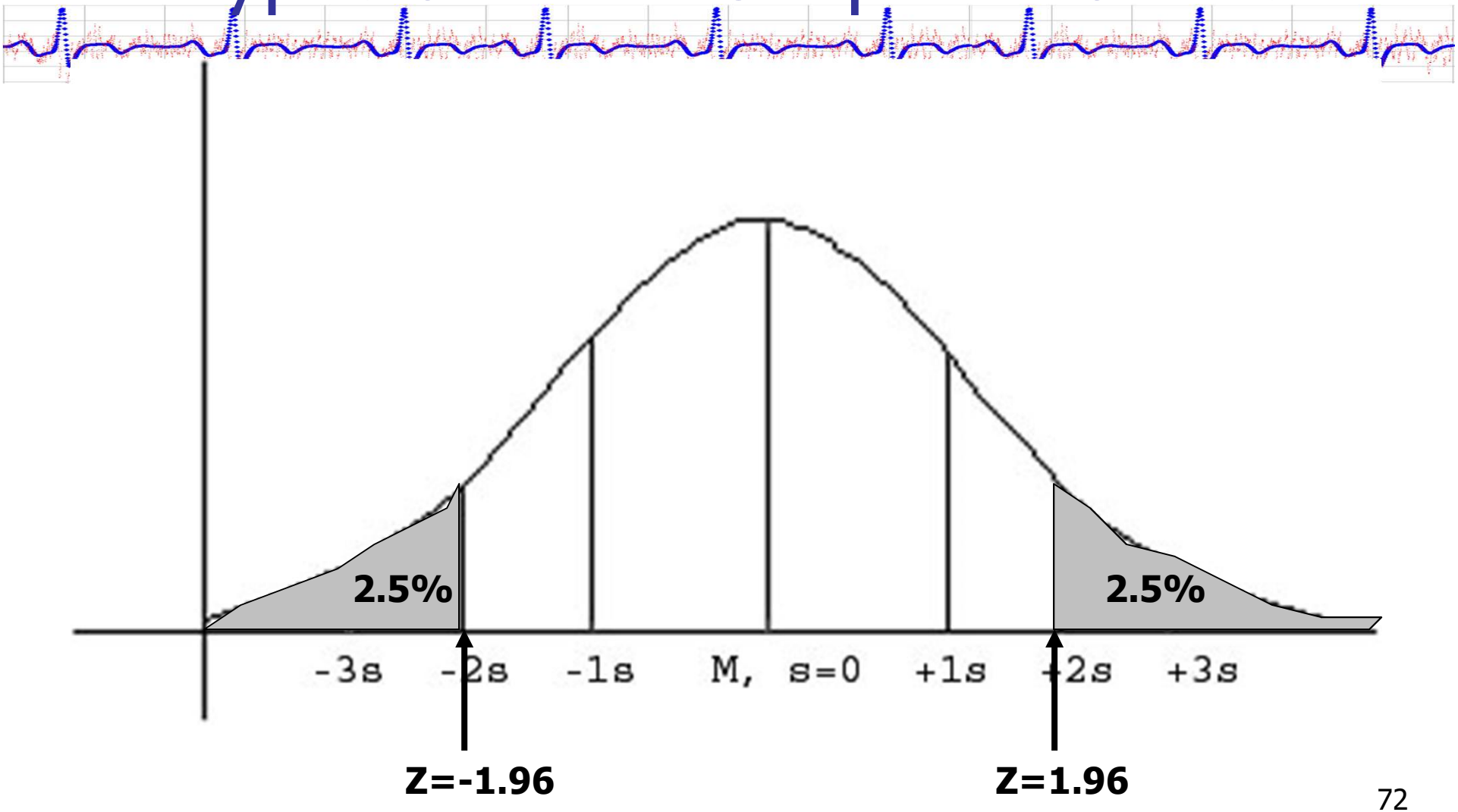
Say M=170

Comparison Distribution = Distribution of Means

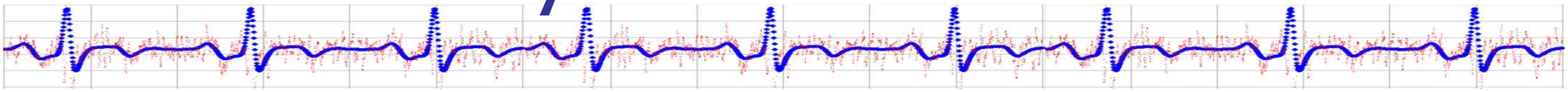


- N = 10
- $\mu = 149$, $\sigma = 12$, $\sigma^2 = 144$
- $\mu_M = 149$, $\sigma_M^2 = 144/10 = 14.4$
- Z = (170 – 149) / $\sqrt{14.4} = 5.5$

Example: 10 admins using WW types at M=170 wpm 2 tail test



Don’t try this at home



- You would never do a study this way.
- Why?
  - Can’t control extraneous variables through randomization.
  - Usually don’t know population statistics.
  - ~~Can’t generalize from an individual.~~