MICROENGINEERING 110:
Design and analysis of experiments – Statistics for Experimenters

Module 3: Comparison Statistics

Prof. Vivek Subramanian

Microengineering

EPFL

Reminder: Probability and Inferential Statistics

- **For any known population it is possible to determine the probability of obtaining any specific sample.**
- **Typically a researcher begins with a sample.**
- **If the sample has a high probability of being obtained from a specific population, then the researcher can conclude that the sample is likely to have come from that population.**
- **If the sample has a very low probability of being obtained from a specific population, then it is reasonable for the researcher to conclude that the specific population is probably not the source for the sample.**

Error and power

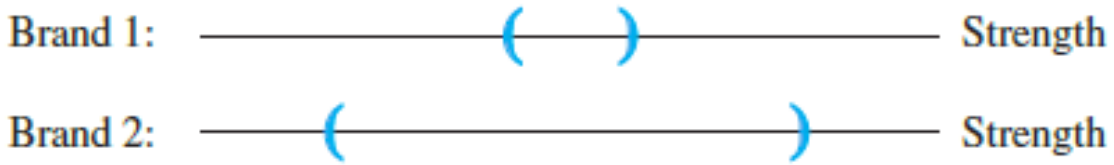
- **Type I error rate (or significance level): the probability of finding an effect that isn't real (false positive).**
  - **If we require p-value<.05 for statistical significance, this means that 1/20 times we will find a positive result just by chance.**
- **Type II error rate: the probability of missing an effect (false negative).**
- **Statistical power: the probability of finding an effect if it is there (the probability of not making a type II error).**

Confidence Intervals

- **We can use a "confidence interval" to study the precision of an estimate.**

- **Example:**

- **Consider I-Beams used to construct buildings**
  - **Brand 1 and Brand 2 happen to have identical yield strengths**
  - **However, when we look at distributions of data, we find that 95% of samples of Brand 1 fall within a tighter range than 95% of samples of Brand 2**



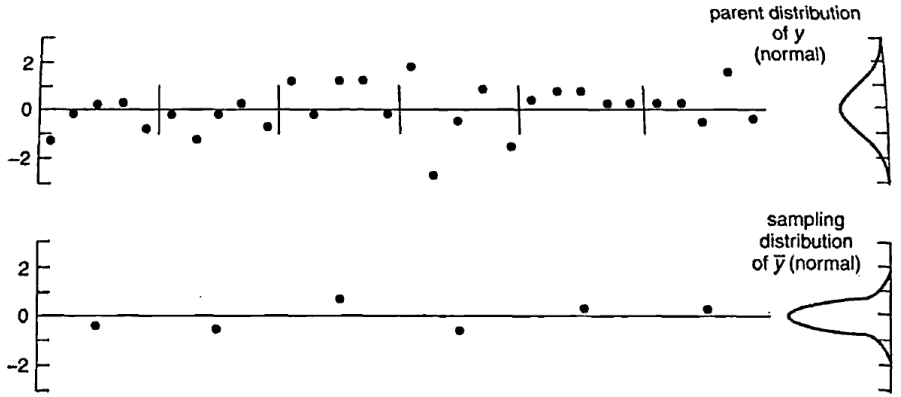
  - **These represent the 95% confidence intervals of Brands 1 an 2**
  - **Which one would you pick, and why? How would this relate to the needs of your construction?**

Sample Size and Confidence Intervals

- **In most experiments, we don't just take 1 sample… we take several samples, and the look at the mean and standard deviation.**
- **If we want to determine our confidence in this result, we are effectively doing "repeated sampling", i.e., we are asking how confident we are that the mean will fall within a certain range, we should follow our repeated sampling methodology from Module 2.**

	Parent Distribution for Observations y	Sampling Distribution for Averages $\bar{y}$
Mean	η	η
Variance	σ^2	σ^2/n
Standard deviation	σ	$\sigma/\sqrt{n}$
Form of parent distribution	Any*	More nearly normal than the parent distribution

*This statement applies to all parent distributions commonly met in practice. It is not true for certain mathematical toys (e.g., the Cauchy distribution), which need not concern us here.



- **So, the resulting normalization for the distribution will be:**

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Confidence Intervals in normal distributions

- **For the normal distribution**

$$P\left(-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = .95$$

- **More generally:**

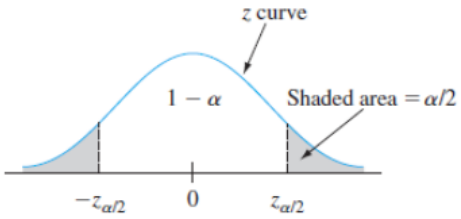


Figure 7.4 $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$

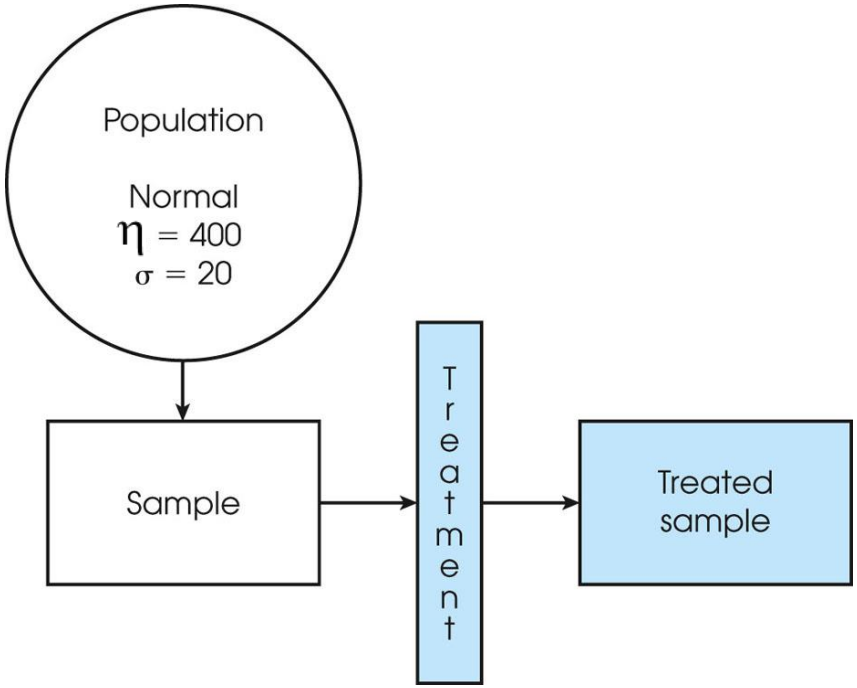
A **100(1 − α)% confidence interval** for the mean μ of a normal population when the value of σ is known is given by

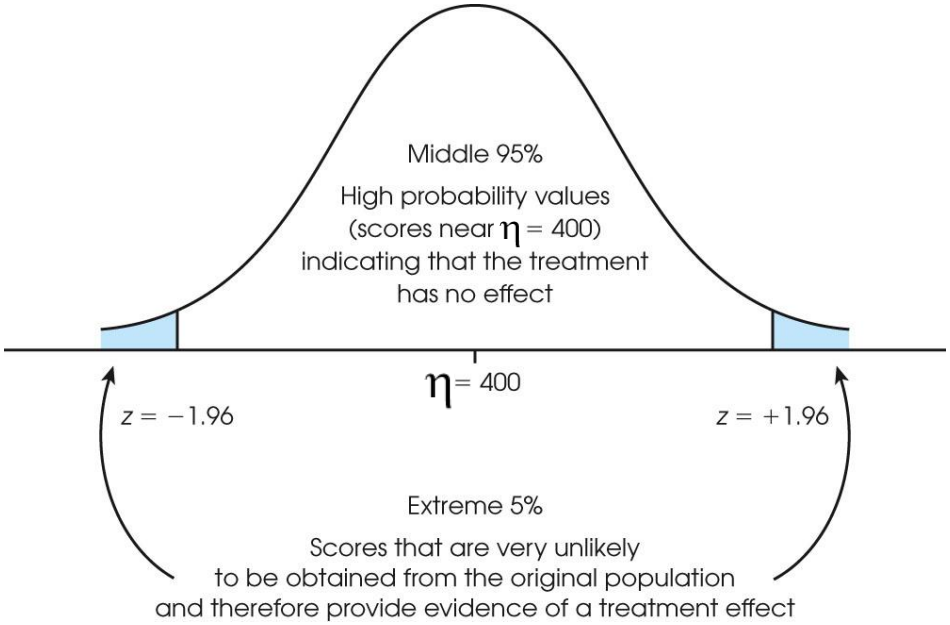
$$\left(\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

or, equivalently, by $\overline{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$.

Confidence Intervals and experiments

We use confidence intervals to check if an experimental factor affects a measured response





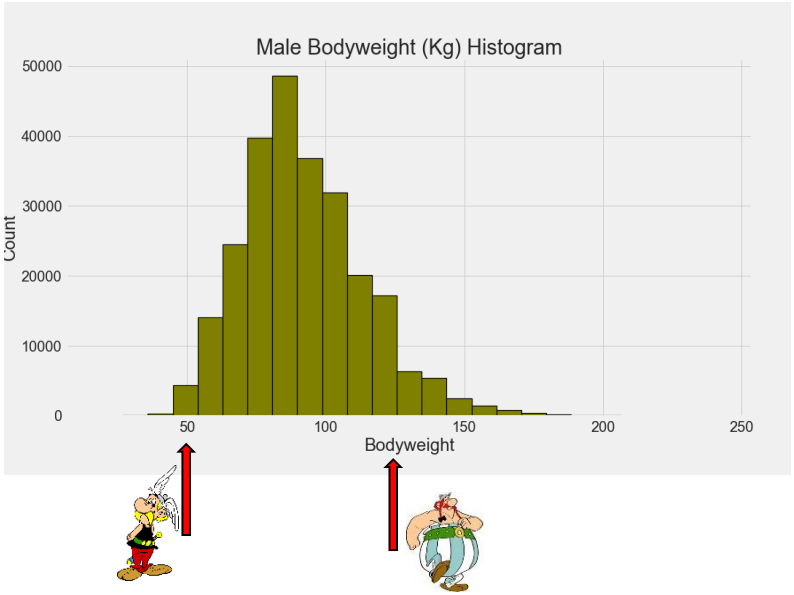
Example: Asterix

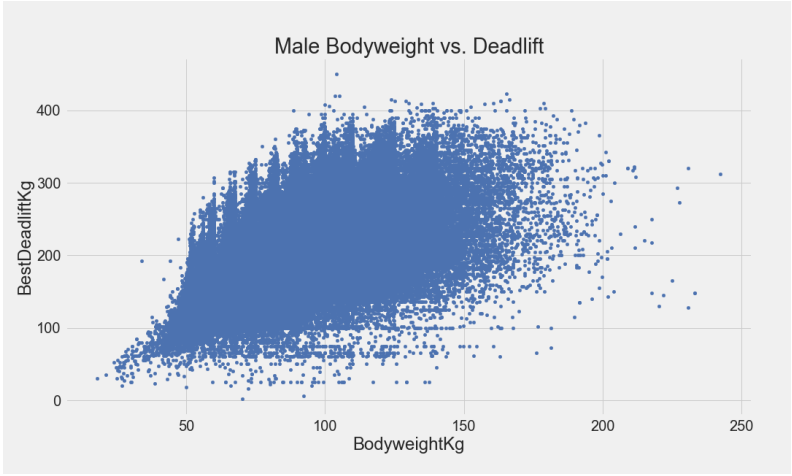
Astérix

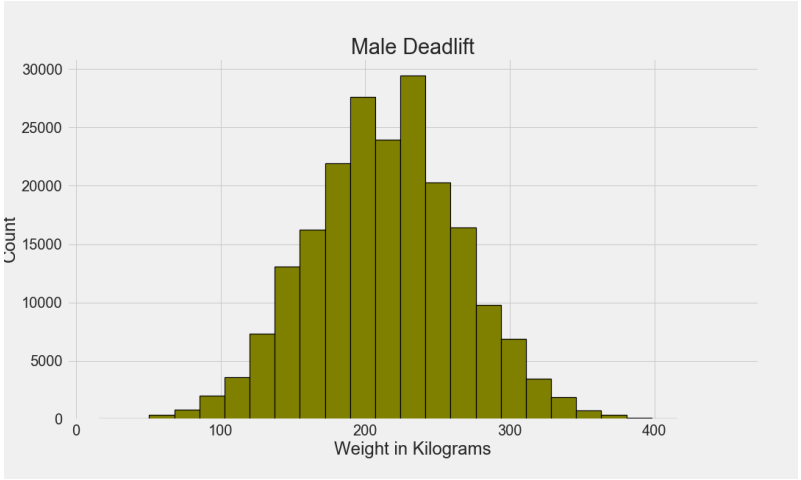
Panoramix's
Potion magique

Obélix

Weightlifting population data



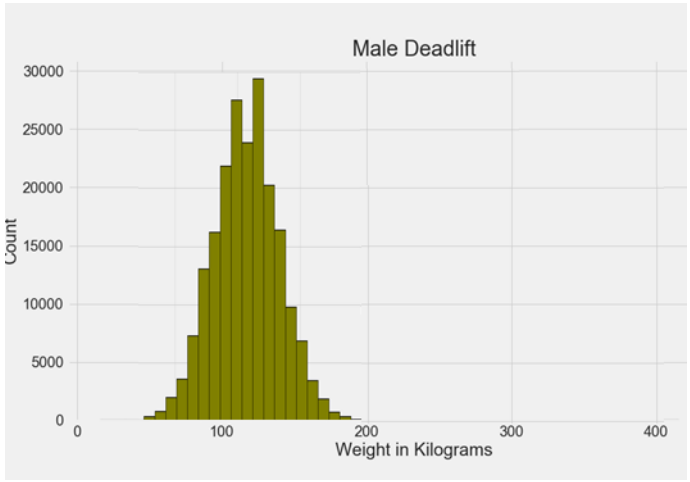




Questions to consider:

- Are body weight and deadlift weight independent?
- Assuming normality, sketch an expected deadlift histogram for an Asterix-like human (pre-potion)
- Repeat for an Obelix-like human (pre-potion)
- What conclusions might you draw regarding bodyweight vs. deadlift weight?

Does the potion work?



Scenario 1:
Pre-potion deadlift: 100kg
Post-potion deadlift: 150kg

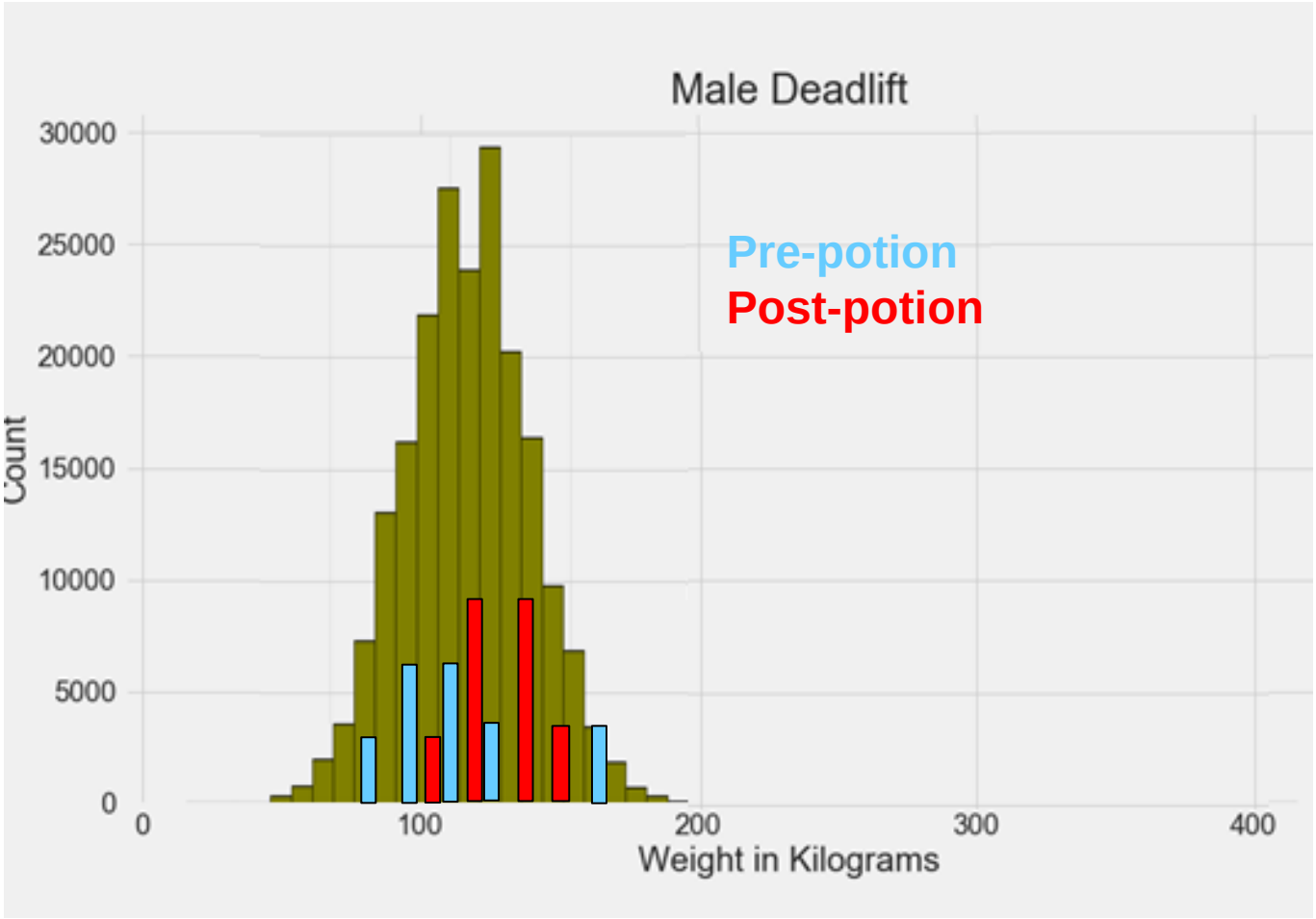
Scenario 2:
Pre-potion deadlift: ?
Post-potion deadlift: 150kg

Scenario 3:
Pre-potion deadlift: ?
Post-potion deadlift: 250kg

This is often the situation in real experiments, since we don't have access to pre- and post- treatment data on an individual sample

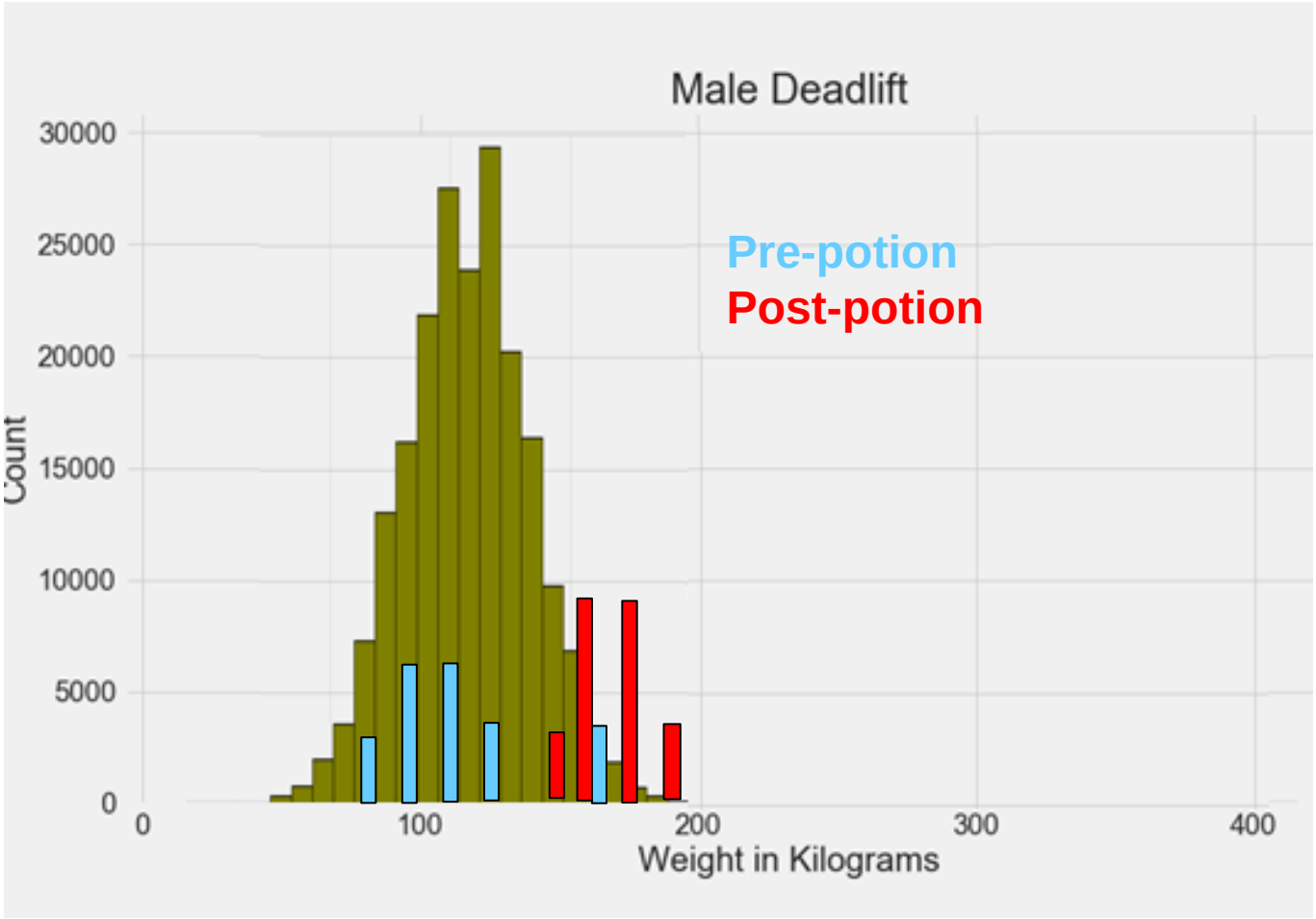
Does the potion work?

Assurancetourix



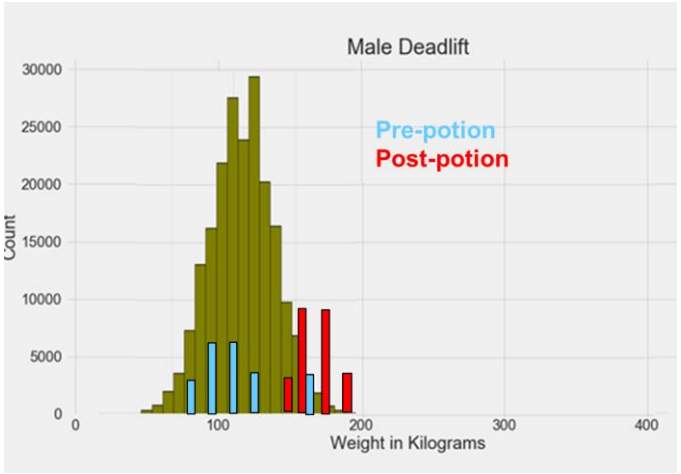
Does the potion work?

Assurancetourix

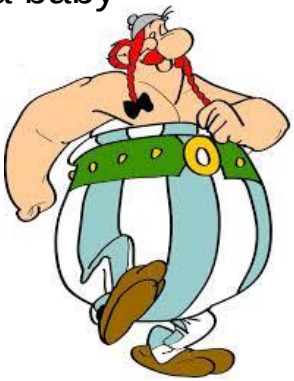


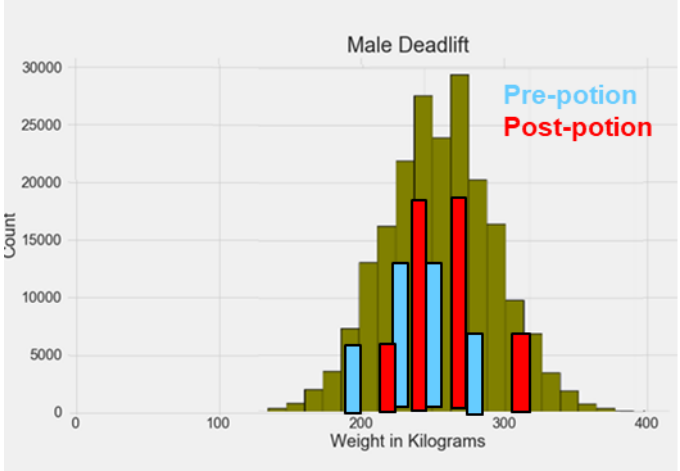
What can you say about the effectiveness of the potion?

Did not fall in potion when he was a baby



Fell in potion when he was a baby





Types of Measures / Variables

- ***Nominal / categorical***
  - **Gender, major, blood type, eye color**
- ***Ordinal***
  - **Rank-order of favorite films**
- ***Interval / scale***
  - **Time, money, age, GPA**

Main Analysis Techniques

Variable Type	Example	Commonly-used Statistical Method
Nominal by Nominal	blood type by gender	Chi-square
Scale by Nominal	GPA by gender GPA by major	t-test Analysis of Variance
Scale by Scale	weight by height GPA by SAT	Regression Correlation

Overview of common statistical tests

Outcome Variable	Are the observations correlated?		Assumptions
	independent	correlated	
Continuous (e.g. blood pressure, age, pain score)	Ttest ANOVA Linear correlation Linear regression	Paired ttest Repeated-measures ANOVA Mixed models/GEE modeling	Outcome is normally distributed (important for small samples). Outcome and predictor have a linear relationship.
Binary or categorical (e.g. breast cancer yes/no)	Chi-square test Relative risks Logistic regression	McNemar's test Conditional logistic regression GEE modeling	Chi-square test assumes sufficient numbers in each cell (>=5)
Time-to-event (e.g. time-to-death, time-to-fracture)	Kaplan-Meier statistics Cox regression	n/a	Cox regression assumes proportional hazards between groups

Continuous outcome (means)

Outcome Variable	Are the observations correlated?		Alternatives if the normality assumption is violated (and small n):
	independent	correlated	
Continuous (e.g. blood pressure, age, pain score)	**Ttest:** compares means between two independent groups **ANOVA:** compares means between more than two independent groups **Pearson's correlation coefficient** (linear correlation): shows linear correlation between two continuous variables **Linear regression:** multivariate regression technique when the outcome is continuous; gives slopes or adjusted means	**Paired ttest:** compares means between two related groups (e.g., the same subjects before and after) **Repeated-measures ANOVA:** compares changes over time in the means of two or more groups (repeated measurements) **Mixed models/GEE modeling**: multivariate regression techniques to compare changes over time between two or more groups	<u>Non-parametric statistics</u> **Wilcoxon sign-rank test:** non-parametric alternative to paired ttest **Wilcoxon sum-rank test** (=Mann-Whitney U test): non-parametric alternative to the ttest **Kruskal-Wallis test:** non-parametric alternative to ANOVA **Spearman rank correlation coefficient:** non-parametric alternative to Pearson's correlation coefficient

A general methodology for hypothesis testing

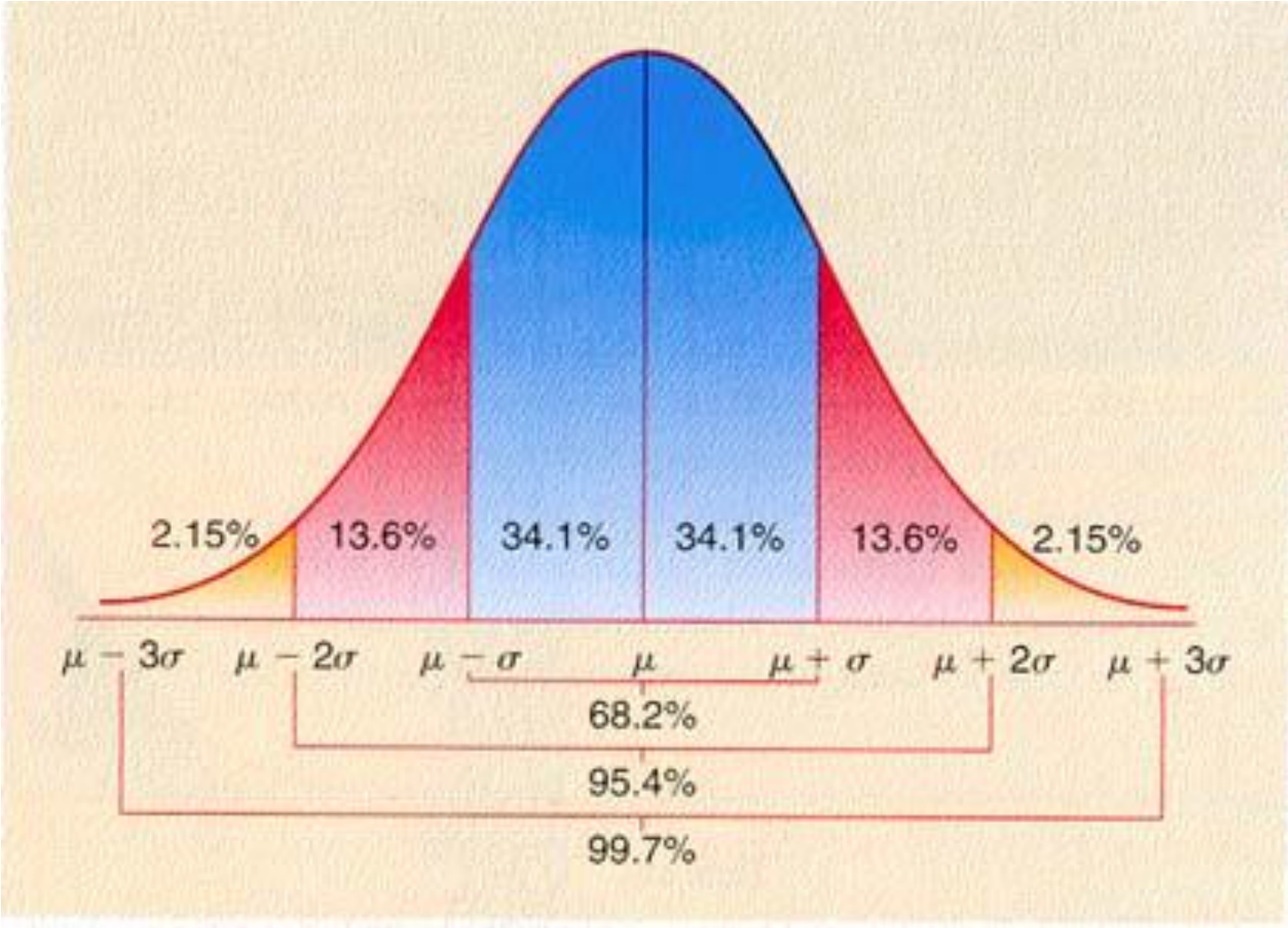
- **Hypothesis to test: The potion works**
- **Null hypothesis: The potion doesn't work, so there is no association between the pre- and post- data, i.e.,**
  - **we would expect that the samples came from the same population**
  - **Any differences in sample means, etc. were purely due to dispersion**

The test:

- **Assuming we have access to the population central tendency and dispersion**
- **Calculate probability that the post-treatment sample came from that population**
  - **If yes: Null hypothesis is probably correct**
  - **If no: Null hypothesis is probably incorrect**

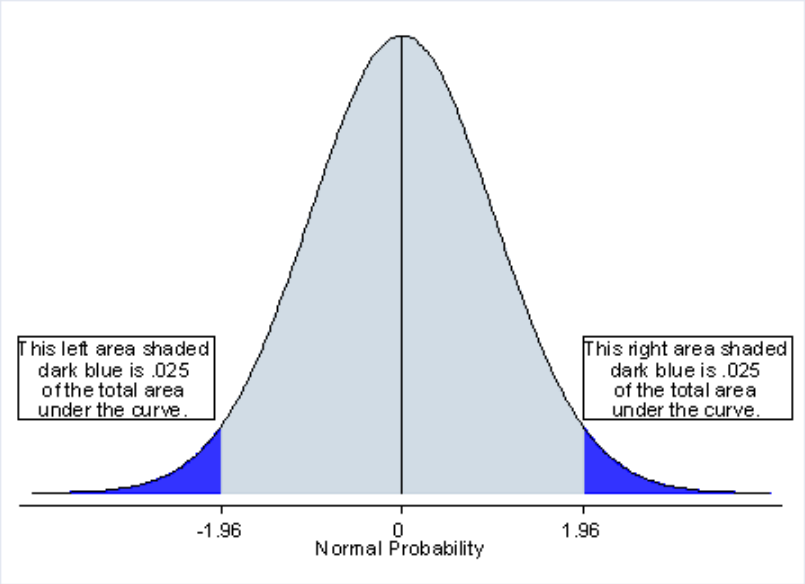
- **What do we do if we don't have the population information?**

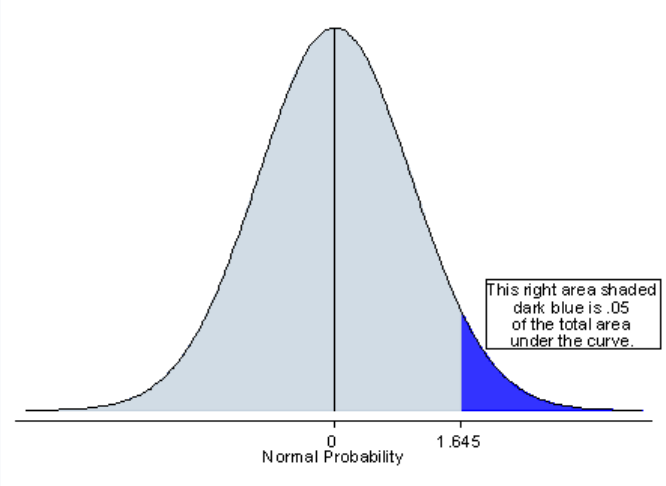
Reminder: Normal Distribution



The Basics

- **Develop an experimental hypothesis**
  - **H0 = null hypothesis**
  - **H1 = alternative hypothesis**
- **Statistically significant result**
  - **P Value = .05**





P-Value

- **Probability that observed result is true**
- **Level = .05 or 5%**
- **95% certain our experimental effect is genuine**

Logic of Statistical Inference

- **What is the *probability* of drawing the observed sample from a universe with no differences?**
- **If probability *very low*, then differences in sample likely reflect differences in universe**
- **Then *null hypothesis* can be rejected; difference in sample is *statistically significant***

Tests that do not make the NIID assumption

- **Empirical past data**
  - **Calculate p value based on comparisons to past data**
- **Randomization test**
  - **Generate a fake data set by assuming the null hypothesis is true, which means that switching treatments should have no impact on the data and we can generate a large data set using computational means**

Z-test

- **Population**

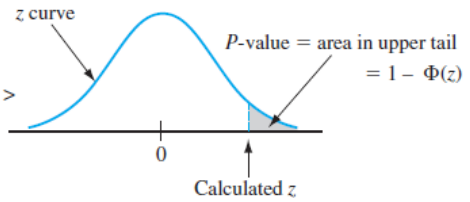
$$z = \frac{\bar{x} - \eta}{\sigma}$$

- **Problems**
  - **Cost**
  - **Not able to include everyone**
  - **Too time consuming**
  - **Ethical right to privacy**

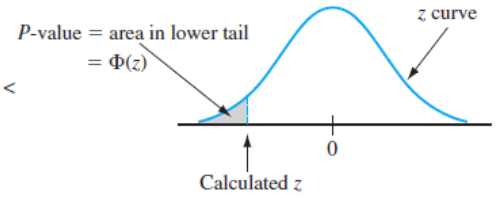
Realistically researchers can only do sample based studies

Z-test and P-values

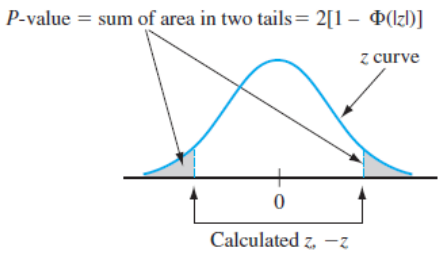
1. **Upper-tailed test**
   H_a contains the inequality >



2. **Lower-tailed test**
   H_a contains the inequality <



3. **Two-tailed test**
   H_a contains the inequality ≠



Null hypothesis: H_0: $\mu = \mu_0$

Test statistic: $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$

Alternative Hypothesis	P-Value Determination
H_a: $\mu > \mu_0$	Area under the standard normal curve to the right of z
H_a: $\mu < \mu_0$	Area under the standard normal curve to the left of z
H_a: $\mu \neq \mu_0$	$2 \cdot$ (area under the standard normal curve to the right of $\lvert z \rvert$)

Assumptions: A normal population distribution with known value of σ.

Note: We find the probability in the shades regions since we would “at least” reach the calculated z value in these regions

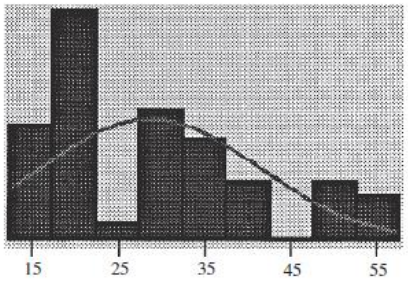
Example: Concrete for sidewalk

- A dynamic cone penetrometer (DCP) is used for measuring material resistance to penetration (in mm) as a cone is driven into sidewalk concrete. Suppose it is required that the true average DCP value less than 30. The concrete will not be used unless there is conclusive evidence that the specification has been met.

- DCP Data:

14.1	14.5	15.5	16.0	16.0	16.7	16.9	17.1	17.5	17.8
17.8	18.1	18.2	18.3	18.3	19.0	19.2	19.4	20.0	20.0
20.8	20.8	21.0	21.5	23.5	27.5	27.5	28.0	28.3	30.0
30.0	31.6	31.7	31.7	32.5	33.5	33.9	35.0	35.0	35.0
36.7	40.0	40.0	41.3	41.7	47.5	50.0	51.0	51.8	54.4
55.0	57.0								

- Generated distribution:



- Does this look normal to you? How should we proceed?

Sidewalk Example (cont'd)

1. μ = true average DCP value
2. H_0: $\mu = 30$
3. H_a: $\mu < 30$ (so the pavement will not be used unless the null hypothesis is rejected)
4. $z = \dfrac{\bar{x} - 30}{s/\sqrt{n}}$
5. With $n = 52$, $\bar{x} = 28.76$, and $s = 12.2647$,

$$z = \frac{28.76 - 30}{12.2647/\sqrt{52}} = \frac{-1.24}{1.701} = -.73$$

6. The P-value for this lower-tailed z test is $\Phi(-.73) = .2327$.

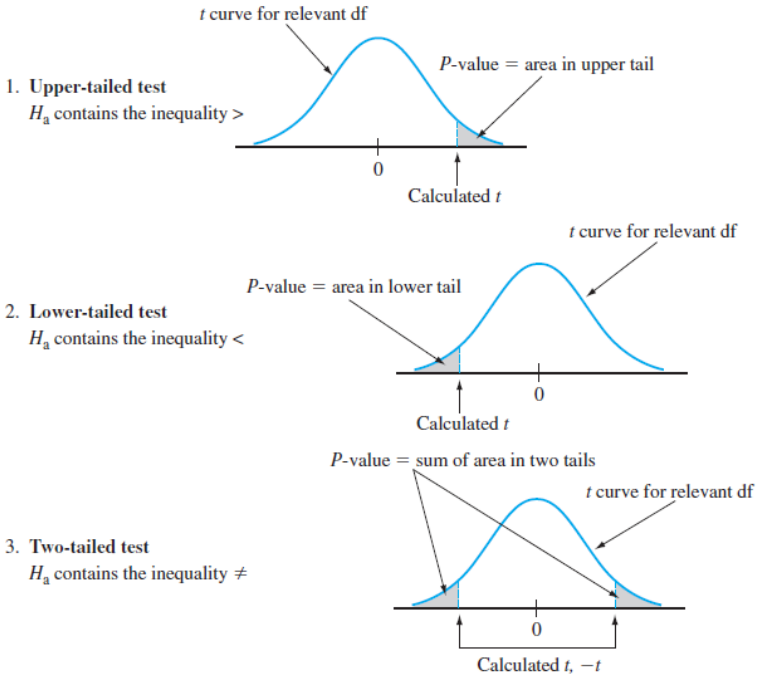
- **So, we don't use the pavement since there is not clear evidence that the concrete will deliver DCP < 30.**

The t-test

- **Used when we don't have access to population standard deviation**

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

- **There are a few different types of t-tests:**
  - **If we know the population mean, and want to see if a sample set is "different", we use a 1-sample t-test**
  - **If we have pre- and post- treatment independent samples, we use a 2-sample t-test**
  - **If the pre- and post-samples are not independent, then we use a paired t-test**



T-test

- **To test a sample vs. a population (i.e., is the sample part of the population)**

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

- **2 Sample t-test**

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{x}_1 - \bar{x}_2}} \qquad s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- **Paired t-test**

1. Calculate the difference ($d_i = y_i - x_i$) between the two observations on each pair, making sure you distinguish between positive and negative differences.
2. Calculate the mean difference, $\bar{d}$.
3. Calculate the standard deviation of the differences, s_d, and use this to calculate the standard error of the mean difference, $SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$
4. Calculate the t-statistic, which is given by $T = \frac{\bar{d}}{SE(\bar{d})}$. Under the null hypothesis, this statistic follows a t-distribution with $n - 1$ degrees of freedom.

Some online p value calculators

- https://www.graphpad.com/quickcalcs/pvalue1.cfm
- https://www.socscistatistics.com/pvalues/
- https://goodcalculators.com/student-t-value-calculator/

An Example: Yield improvement

- **Here is pre- and post-change data from a process that is modified with the goal of improving yield**

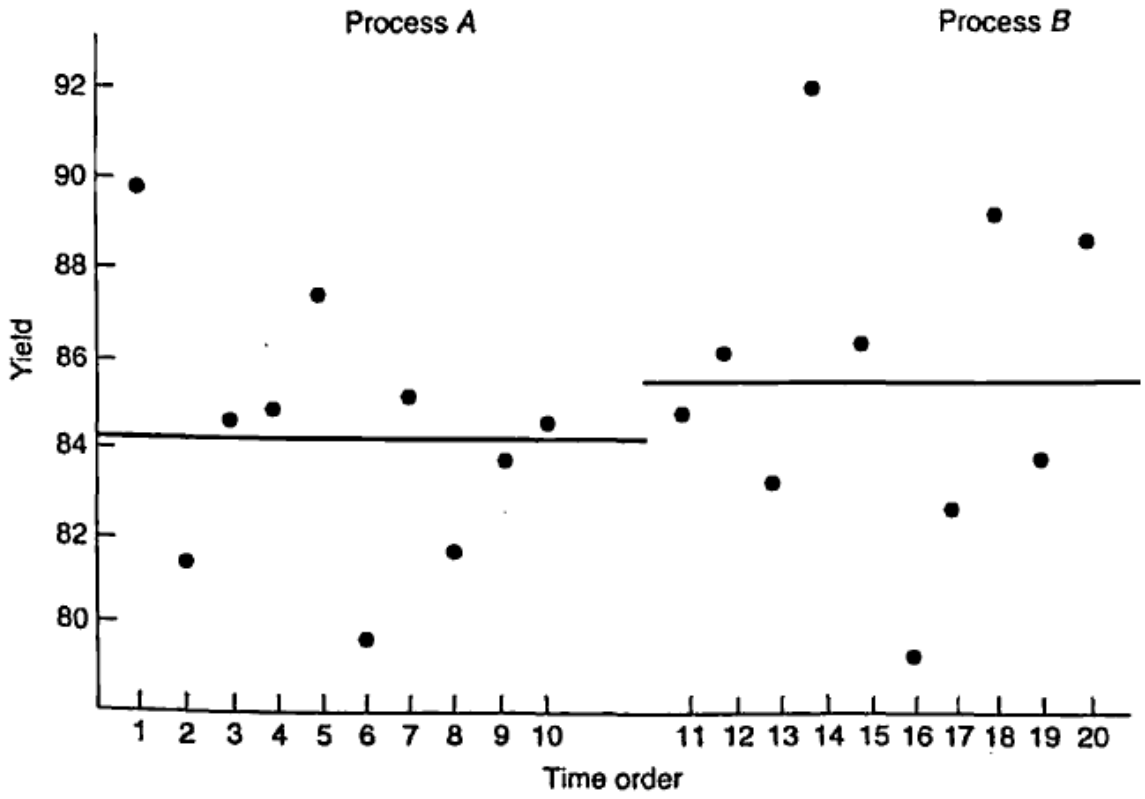


Table 3.1. Yield Data from an Industrial Experiment

Time Order	Method	Yield
1	*A*	89.7
2	*A*	81.4
3	*A*	84.5
4	*A*	84.8
5	*A*	87.3
6	*A*	79.7
7	*A*	85.1
8	*A*	81.7
9	*A*	83.7
10	*A*	84.5
11	*B*	84.7
12	*B*	86.1
13	*B*	83.2
14	*B*	91.9
15	*B*	86.3
16	*B*	79.3
17	*B*	82.6
18	*B*	89.1
19	*B*	83.7
20	*B*	88.5

$\bar{y}_A = 84.24 \qquad \bar{y}_B = 85.54$

$\bar{y}_A - \bar{y}_B = 1.30$

- **Do you think the process improvement works? What type of t-test should we use?**

2 sample t-test

- We have:

$$s_A^2 = \frac{\sum (y_A - \bar{y}_A)^2}{n_A - 1} = \frac{75.784}{9} = 8.42$$

$$s_B^2 = \frac{\sum (y_B - \bar{y}_B)^2}{n_B - 1} = \frac{119.924}{9} = 13.32$$

- Since we assumed similar variances, we can combine these as:

$$s^2 = \frac{\sum (y_A - \bar{y}_A)^2 + \sum (y_B - \bar{y}_B)^2}{n_A + n_B - 2} = \frac{75.784 + 119.924}{18} = 10.87$$

- Next, we can write the t distribution (replace the σ with s in the z distribution):

$$t = \frac{(\bar{y}_B - \bar{y}_A)}{s\sqrt{1/n_B + 1/n_A}}$$

- We also have $\overline{y_B} - \overline{y_A} = 1.30$ and $s\sqrt{1/n_B + 1/n_A}$ =1.47, so t = 0.88, so $\Pr(t > 0.88) = 19.5\%$.

Table 3.1. Yield Data from an Industrial Experiment

Time Order	Method	Yield
1	A	89.7
2	A	81.4
3	A	84.5
4	A	84.8
5	A	87.3
6	A	79.7
7	A	85.1
8	A	81.7
9	A	83.7
10	A	84.5
11	B	84.7
12	B	86.1
13	B	83.2
14	B	91.9
15	B	86.3
16	B	79.3
17	B	82.6
18	B	89.1
19	B	83.7
20	B	88.5

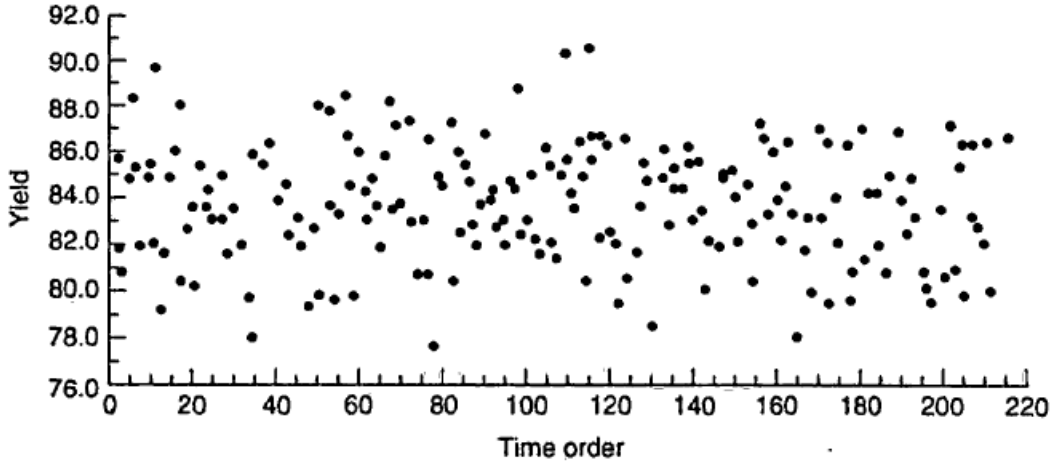
$\bar{y}_A = 84.24 \quad \bar{y}_B = 85.54$

$\bar{y}_A - \bar{y}_B = 1.30$

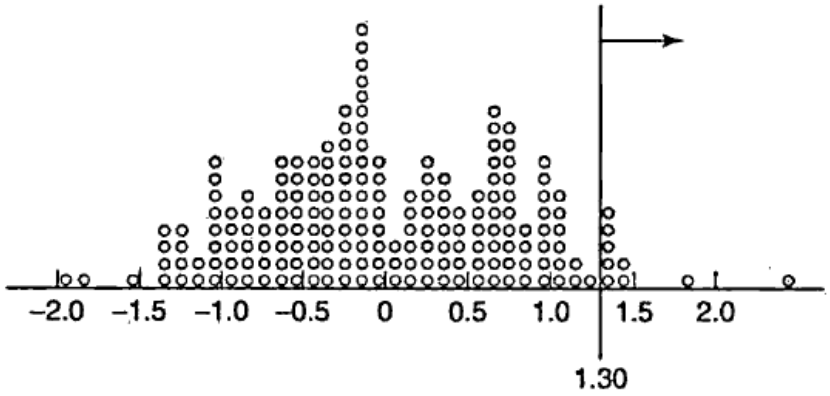
Therefore, process B is not a statistical improvement on process A

But what if I have access to the population data

- **Yield over the last approx. 200 runs**



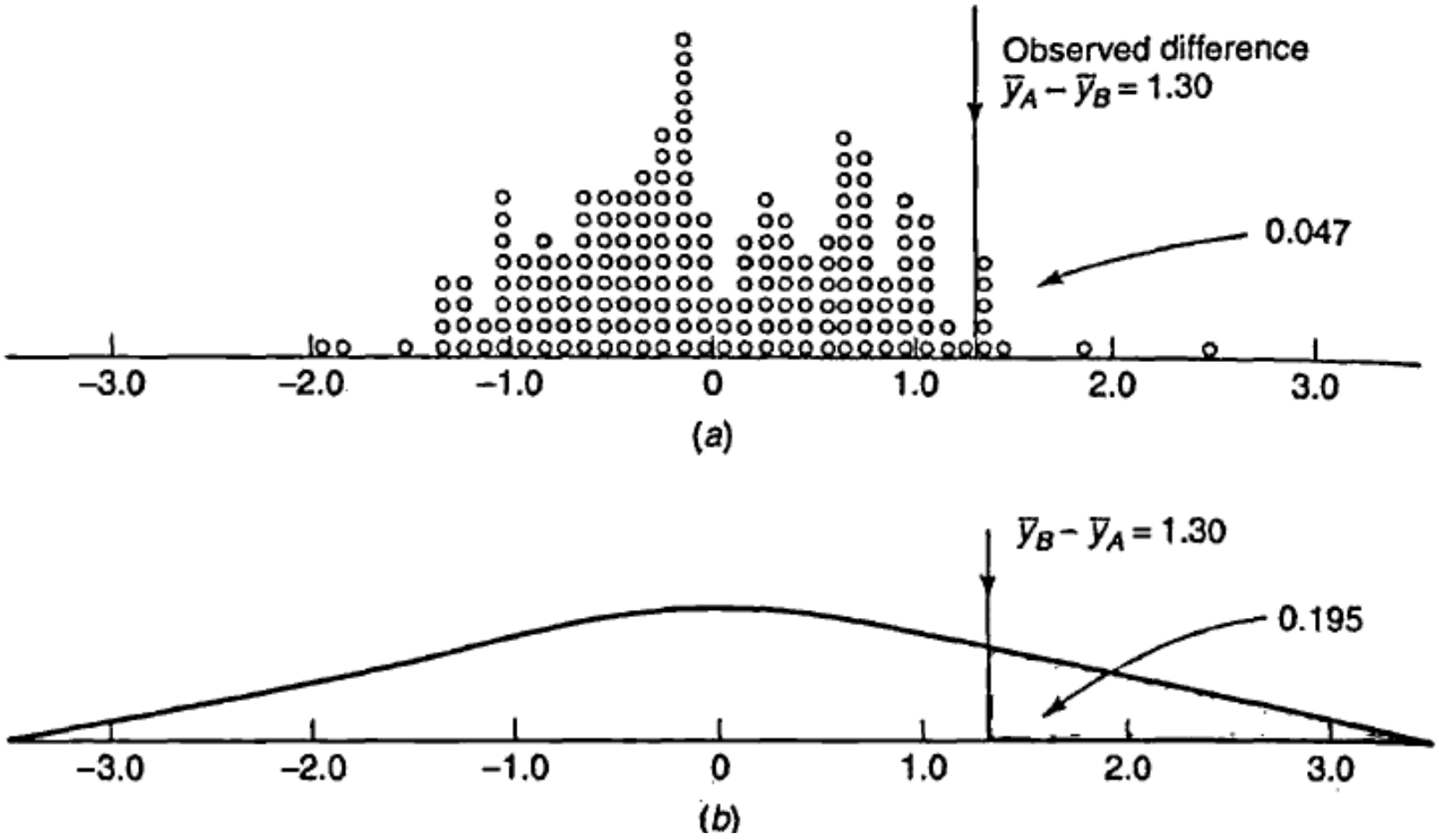
- **I generate a relative reference set of data by plotting the change in $\bar{y}$ for adjacent set of 10 observations**



Only 9/147 observations support the null hypothesis (4.7% significance level), so the process probably works.

Note: to generate this reference set, I needed a lot of prior data, which isn't always available

Comparing the external reference with the t test



The difference is due to breakdown in NIID assumptions.

So what do we do in that case?

- The t-test starts producing errors if the NIID (normal, independent and identically distributed) assumption breaks
- We often don't have external reference data.
- What do we do? Answer: we can test for significant by creating a fake set of random data… this is known as the randomization test.
- Example: Consider the test on a possible "improved" tomato fertilizer

Table 3.3. Results from a Randomized Experiment (Tomato Yields in Pounds)

Position in row	1	2	3	4	5	6	7	8	9	10	11
Fertilizer	*A*	*A*	*B*	*B*	*A*	*B*	*B*	*B*	*A*	*A*	*B*
Pounds of tomatoes	29.2	11.4	26.6	23.7	25.3	28.5	14.2	17.9	16.5	21.1	24.3

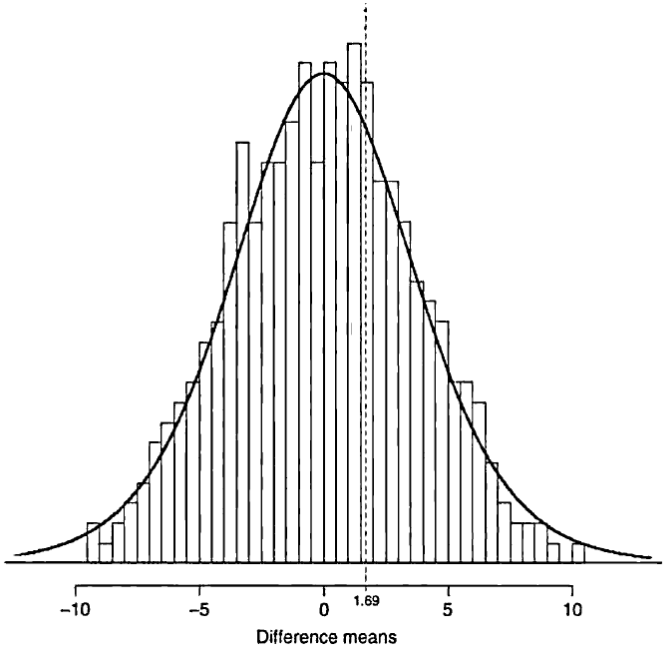
Standard Fertilizer *A*	Modified Fertilizer *B*
29.9	26.6
11.4	23.7
25.3	28.5
16.5	14.2
21.1	17.9
	24.3
$n_A = 5$	$n_B = 6$
$\bar{y}_A = 20.84$	$\bar{y}_B = 22.53$

Randomization test

- If the null hypothesis is true, we should be able to mix up the A and B labels and it shouldn't change the population
- In the actual data, there are 5 "A" fertilizers and 6 "B" fertilizers. Therefore, if were to randomize the labels, there would be a total of $\frac{11!}{5!6!} = 462$ variations

Position in row	1	2	3	4	5	6	7	8	9	10	11
Fertilizer	*A*	*A*	*B*	*B*	*A*	*B*	*B*	*B*	*A*	*A*	*B*
Pounds of tomatoes	29.2	11.4	26.6	23.7	25.3	28.5	14.2	17.9	16.5	21.1	24.3

- These can then be used plot the distribution of faked $\overline{y_B} - \overline{y_A}$ results.
- Recall the real $\overline{y_B} - \overline{y_A}$ = 1.69
- We see that $154/462 = 33\%$. equal or exceed that value, which means that the fertilizer improvement likely meets the null hypothesis, i.e., it doesn't actually improve anything!



t-test on the tomato experiment

$$\bar{y}_B - \bar{y}_A = 22.53 - 20.84 = 1.69$$

$$s_A^2 = \frac{\sum y_A^2 - \left(\sum y_A\right)^2 / n_A}{n_A - 1} = \frac{S_A}{\nu_A} = \frac{209.9920}{4} = 52.50$$

$$s_B^2 = \frac{\sum y_B^2 - \left(\sum y_B\right)^2 / n_B}{n_B - 1} = \frac{S_B}{\nu_B} = \frac{147.5333}{4} = 29.51$$

The pooled variance estimate is

$$s^2 = \frac{S_A + S_B}{\nu_A + \nu_B} = \frac{\nu_A s_A^2 + \nu_B s_B^2}{\nu_A + \nu_B} = \frac{4(52.50) + 5(29.51)}{4 + 5} = 39.73$$

with $\nu = n_A + n_B - 2 = \nu_A + \nu_B = 9$ degrees of freedom

The estimated variance of $\bar{y}_B - \bar{y}_A$ is $s^2(1/n_B + 1/n_A) = 39.73(1/6 + 1/5) = 14.57$.
The standard error of $\bar{y}_B - \bar{y}_A$ is $\sqrt{14.57} = 3.82$.

$$t_0 = \frac{(\bar{y}_B - \bar{y}_S)}{\sqrt{s^2(1/n_B + 1/n_A)}} = \frac{(22.53 - 20.84)}{\sqrt{39.73(1/6 + 1/5)}} = \frac{1.69}{3.82}$$

$$t_0 = 0.44 \qquad \text{with} \qquad \nu = 9 \text{ degrees of freedom}$$

$$\Pr(t \geq t_0) = \Pr(t \geq 0.44) = 0.34$$

Similar to randomization experiment, so the fertilizer doesn't offer any improvement

Paired Experiments

- **Many experiments result in pairs (e.g., we have before and after data on a specific sample, or we ran two “identical” samples through highly correlated processes other than the treatment in question)**

paired t test

1. Calculate the difference ($d_i = y_i - x_i$) between the two observations on each pair, making sure you distinguish between positive and negative differences.
2. Calculate the mean difference, $\bar{d}$.
3. Calculate the standard deviation of the differences, s_d, and use this to calculate the standard error of the mean difference, $SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$
4. Calculate the t-statistic, which is given by $T = \frac{\bar{d}}{SE(\bar{d})}$. Under the null hypothesis, this statistic follows a t-distribution with $n - 1$ degrees of freedom.
5. Use tables of the t-distribution to compare your value for T to the t_{n-1} distribution. This will give the p-value for the paired t-test.

Compare to the normal t test:

$$t = \frac{differences\_between\_sample\_means}{estimated\_standard\_error\_of\_differences\_between\_means}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

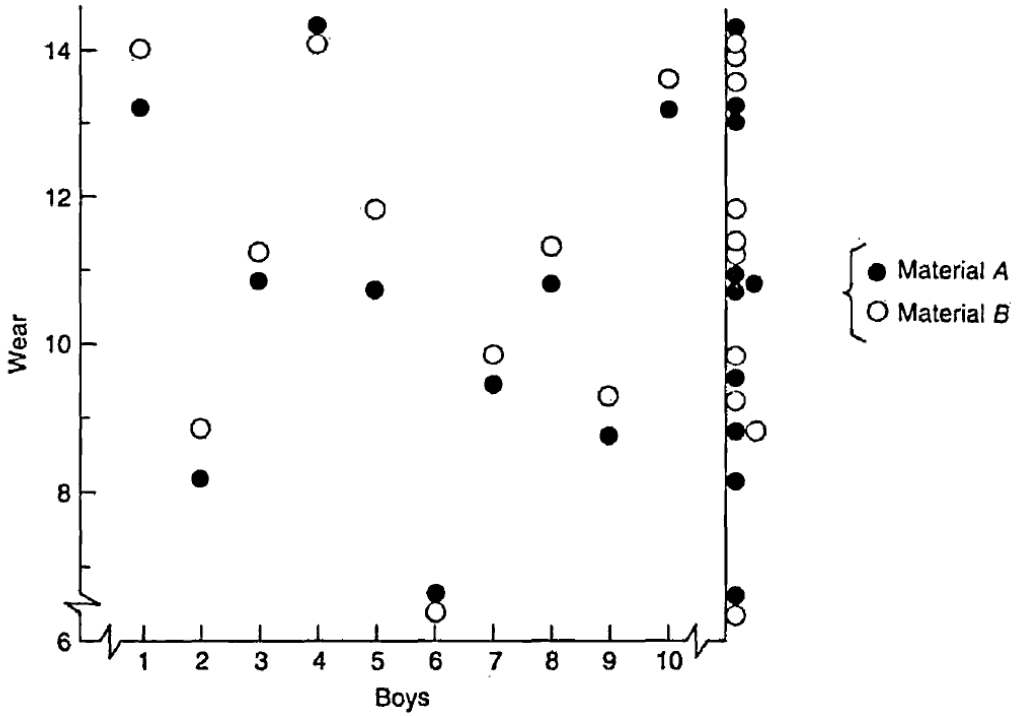
$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Example: An improved shoe sole material?

- Put a shoe with the old sole material on one foot of each boy
- Put a shoe with the new sole material on the other foot
- The soles are now paired, under the assumption that they will see similar conditions
- Left and Right foot assignments are randomly assigned using a coin toss
- Measure the wear on the soles for 10 boys

Table 3.5. Boys' Shoes Example: Data on the Wear of Shoe Soles Made of Two Different Materials *A* and *B*

Boy	Material *A*	Material *B*	Difference $d = B - A$
1	13.2(L)	14.0(R)	0.8
2	8.2(L)	8.8(R)	0.6
3	10.9(R)	11.2(L)	0.3
4	14.3(L)	14.2(R)	−0.1
5	10.7(R)	11.8(L)	1.1
6	6.6(L)	6.4(R)	−0.2
7	9.5(L)	9.8(R)	0.3
8	10.8(L)	11.3(R)	0.5
9	8.8(R)	9.3(L)	0.5
10	13.3(L)	13.6(R)	0.3
	Average difference $\bar{d} = 0.41$		



Paired t-test

- **The sample variance is:**

$$s_d^2 = \frac{\sum (d - \bar{d})^2}{n - 1} = 0.149$$

- **Assuming random sampling, we can say:**

$$s_d = \sqrt{0.149} = 0.386 \qquad \text{and} \qquad s_{\bar{d}} = \frac{s_d}{\sqrt{n}} = \frac{0.386}{\sqrt{10}} = 0.122$$

- **For the null hypothesis:**

$$t_0 = \frac{\bar{d} - \delta_0}{s_{\bar{d}}} = \frac{0.41 - 0}{0.12} = 3.4 \qquad \Pr(t \geq 3.4) \cong 0.4\%$$

- **So, we conclude that B wears faster than A**

Randomization test on the shoe wear data

- **Coin Toss**
  - **Heads: Material A on right foot**
  - **Tails: Material B on right foot**
- **Actual coin toss**

T T H T H T T T H T

- **Total number of toss possibilities: 2^{10}=1024**

- **If the null hypothesis is true, the actual choice of H or T shouldn't matter**

$$\bar{d} = \frac{\pm 0.8 \pm 0.6 \pm \cdots \pm 0.3}{10}$$

Results

- **The randomized distribution vs the actual sample**

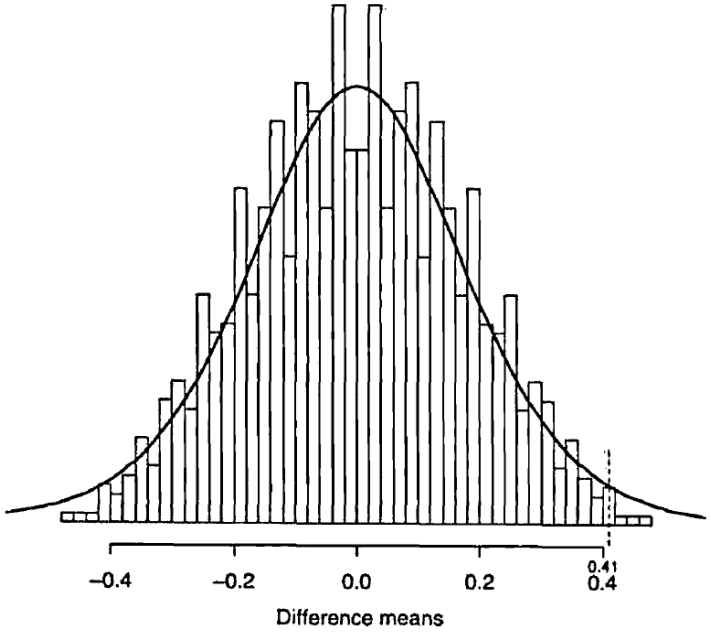


Table 3.5. Boys' Shoes Example: Data on the Wear of Shoe Soles Made of Two Different Materials *A* and *B*

Boy	Material *A*	Material *B*	Difference $d = B - A$
1	13.2(L)	14.0(R)	0.8
2	8.2(L)	8.8(R)	0.6
3	10.9(R)	11.2(L)	0.3
4	14.3(L)	14.2(R)	−0.1
5	10.7(R)	11.8(L)	1.1
6	6.6(L)	6.4(R)	−0.2
7	9.5(L)	9.8(R)	0.3
8	10.8(L)	11.3(R)	0.5
9	8.8(R)	9.3(L)	0.5
10	13.3(L)	13.6(R)	0.3

Average difference $\bar{d} = 0.41$

- **0.41 is actually extremely unlikely in the null hypothesis (5/1024 = 0.5%)**

- **Therefore, we conclude that the soles are actually different and material A is better than B (i.e., B has more wear than A)**

A note about 1-sided vs. 2-sided tests

- **All our probabilities have looked at the one-sided values, since we have assumed that the modification would either improve or degrade the same (or do nothing, i.e., the null hypothesis)**
- **If your modification can do either, it is appropriate to use the two-sided probability tables, which account for both sides of the distribution**

ANOVA: Analysis of Variance

- **The difference between ANOVA and the t tests is that ANOVA can be used in situations where there are *two or more* means being compared, whereas the t tests are limited to situations where only two means are involved.**

- **Analysis of variance is necessary to protect researchers from excessive risk of a Type I error in situations where a study is comparing more than two population means.**

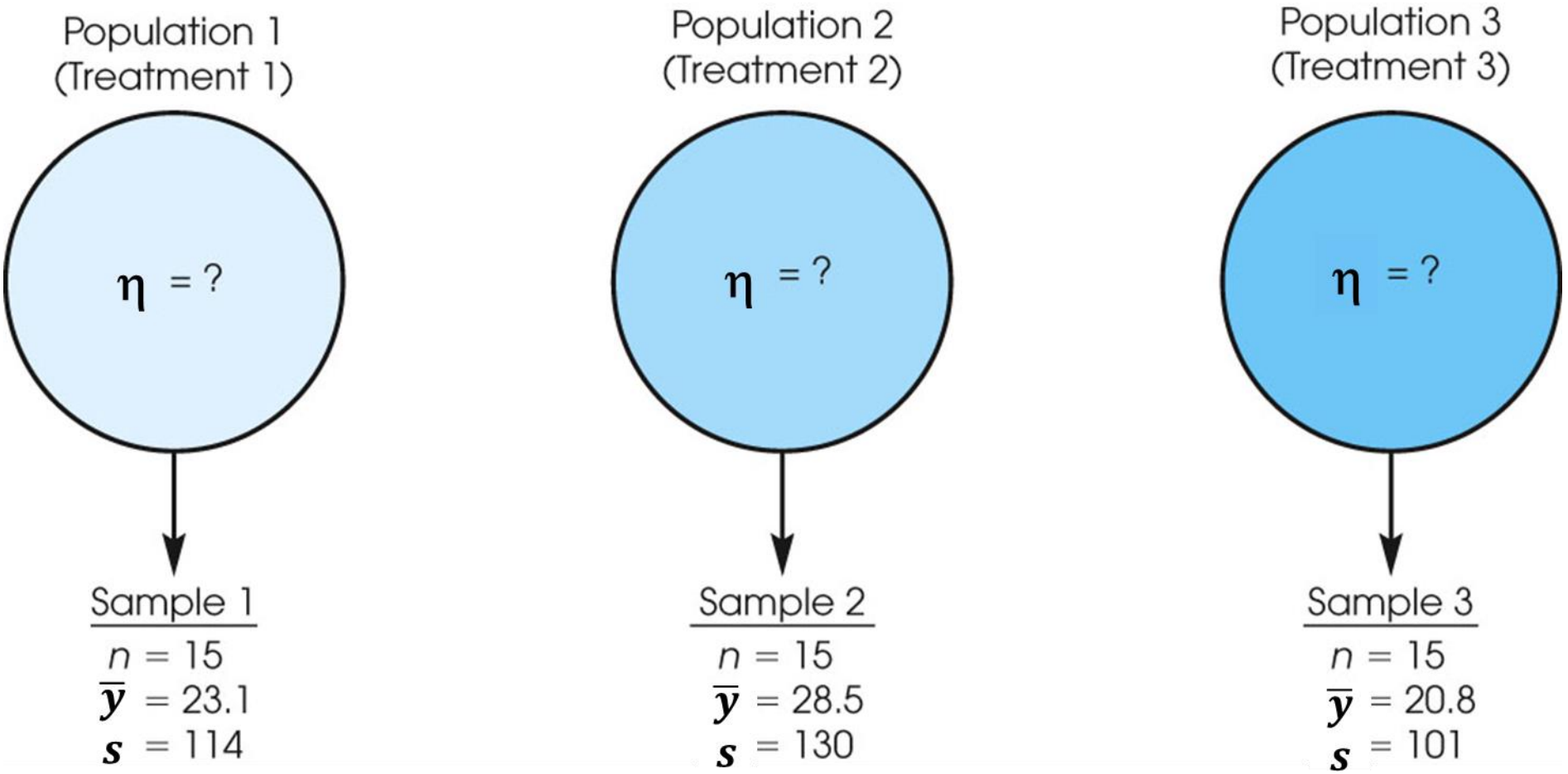
- **ANOVA allows researcher to evaluate all of the mean differences in a single hypothesis test, keeps the risk of a Type I error under control no matter how many different means are being compared.**

Why not use multiple t tests?

- These situations would require a series of several t tests to evaluate all of the mean differences. (Remember, a t test can compare only 2 means at a time.)

- Errors can also accumulate over the multiple t tests

How would we solve this problem using a t-test?



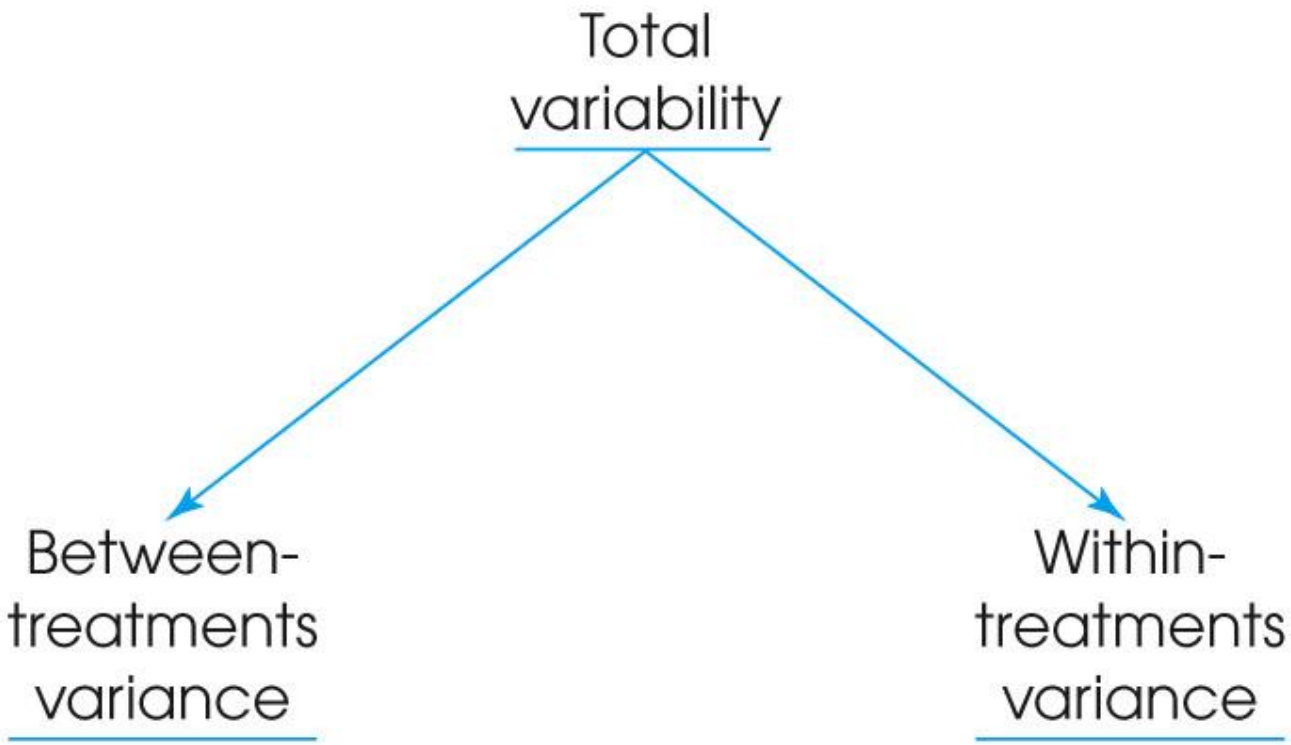
ANOVA methodology

- **The test statistic for ANOVA is an F-ratio, which is a ratio of two sample variances. In the context of ANOVA, the sample variances are called mean squares, or MS values.**
- **The top of the F-ratio $MS_{between}$ measures the size of mean differences between samples. The bottom of the ratio MS_{within} measures the magnitude of differences that would be expected without any treatment effects.**

$$F = \frac{\text{obtained mean differences (including treatment effects)}}{\text{differences expected by chance (without treatment effects)}} = \frac{MS_{between}}{MS_{within}}$$

- **A large value for the F-ratio indicates that the obtained sample mean differences are greater than would be expected if the treatments had no effect.**

Why are these values important?

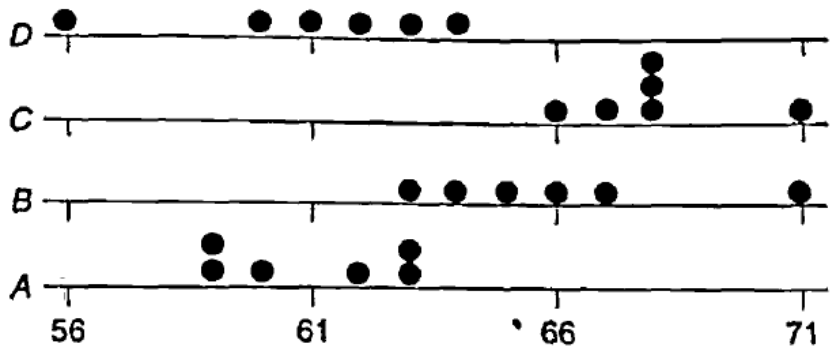


Measures differences due to
1. Systematic treatment effects
2. Random, unsystematic factors

Measures differences due to
1. Random, unsystematic factors

ANOVA Example: Blood coagulation based on diet

- **We consider the effect of 4 different diets on blood coagulation time**



	Diets (Treatments)			
	A	*B*	*C*	*D*
	62	63	68	56
	60	67	66	62
	63	71	71	60
	59	64	67	61
	63	65	68	63
	59	66	68	64
Treatment average	61	66	68	61
Grand average	64	64	64	64
Difference	−3	+2	+4	−3

ANOVA table generation

- **We create the following table**

	Diets (Treatments)			
	A	*B*	*C*	*D*
Treatment average	61	66	68	61
Grand average	64	64	64	64
Difference	−3	+2	+4	−3

Observations	Deviations from Grand Average of 64	Treatment Deviations	Residuals within-Treatment Deviations
y_{ti}	$y_{ti} - \bar{y}$	$\bar{y}_t - \bar{y}$	$y_{ti} - \bar{y}_t$
62 63 68 56	−2 −1 4 −8	−3 2 4 −3	1 −3 0 −5
60 67 66 62	−4 3 2 −2	−3 2 4 −3	−1 1 −2 1
63 71 71 60	−1 7 7 −4	−3 2 4 −3	2 5 3 −1
59 64 67 61	−5 0 3 −3	−3 2 4 −3	−2 −2 −1 0
63 65 68 63	−1 1 4 −1	−3 2 4 −3	2 −1 0 2
59 66 68 64	5 2 4 0	−3 2 4 −3	−2 0 0 3
Y	**D = Y − 64** =	**T** +	**R**
Sum of squares	340 =	228 +	112
degrees of freedom	23 =	3 +	20

Explained next

Reminder: Degrees of freedom

- Degrees of freedom are important in ANOVA
- For example, if we fix the overall mean, the residuals from the overall mean have 23 degrees of freedom since there are 24 data points, and knowing 23 allows us to determine the last 1 from the overall mean
- So, we see that:
  - $y_{ti} - \bar{y}$ has 23 degrees of freedom (since there are 24 residuals from the overall mean
  - $\bar{y}_t - \bar{y}$ has 3 degrees of freedom (since there are 4 treatment mean residuals from the overall mean)
  - $y_{ti} - \bar{y}_t$ has 20 degrees of freedom (there are 24 values, but columns must total to zero, and the overall sum most be zero, so 24-1-3)

The sum of squares

Observations	Deviations from Grand Average of 64		Treatment Deviations		Residuals within-Treatment Deviations	
y_{ti}	$y_{ti} - \bar{y}$		$\bar{y}_t - \bar{y}$		$y_{ti} - \bar{y}_t$	
62 63 68 56	−2 −1 4 −8		−3 2 4 −3		1 −3 0 −5	
60 67 66 62	−4 3 2 −2		−3 2 4 −3		−1 1 −2 1	
63 71 71 60	−1 7 7 −4		−3 2 4 −3		2 5 3 −1	
59 64 67 61	−5 0 3 −3		−3 2 4 −3		−2 −2 −1 0	
63 65 68 63	−1 1 4 −1		−3 2 4 −3		2 −1 0 2	
59 66 68 64	5 2 4 0		−3 2 4 −3		−2 0 0 3	
Y	**D = Y − 64**	=	**T**	+	**R**	
Sum of squares	340	=	228	+	112	
degrees of freedom	23	=	3	+	20	

$$S_D = (-2)^2 + (-1)^2 + (4)^2 + \cdots + (0)^2 = 340$$

$$S_T = (-3)^2 + (2)^2 + (4)^2 + \cdots + (-3)^2 = 228$$

$$S_R = (1)^2 + (-3)^2 + (0)^2 + \cdots + (3)^2 = 112$$

Mean squares and F value

- **We have:**

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	
Between treatments	$S_T = 228$	$v_T = 3$	$m_T = 76.0$	$F_{3,20} = 13.63$
Within treatments	$S_R = 112$	$v_R = 20$	$m_R = 5.6$	
Total about the grand average	$S_D = 340$	$v_D = 23$		

- **Where S_T and S_R and v_T and v_R have been calculated as before**
- **$m_T = S_T / v_T$ and $m_R = S_R / v_R$**
- **$F = m_T / m_R$**
- **Looking up tables for $F_{3,20}$, we find:** $F_{3,20} \geq 13.6$ is less than 0.001.

- **In other words, the null hypothesis is strongly rejected**

How to interpret F-ratio values

- **When the null hypothesis is true and there are no differences between treatments, the F-ratio is balanced.**

- **That is, when the "treatment effect" is zero, the top and bottom of the F-ratio are measuring the same variance.**

- **In this case, you should expect an F-ratio near 1.00. When the sample data produce an F-ratio near 1.00, we will conclude that there is no significant treatment effect.**

- **On the other hand, a large treatment effect will produce a large value for the F-ratio. Thus, when the sample data produce a large F-ratio we will reject the null hypothesis and conclude that there are significant differences between treatments.**

Analysis of Variance and Post Tests

- The null hypothesis for ANOVA states that for the general population there are no mean differences among the treatments being compared; H_0: $\eta_1 = \eta_2 = \eta_3 = \ldots$
- When the null hypothesis is rejected, the conclusion is that there are significant mean differences.
- However, the ANOVA simply establishes that differences exist, it does not indicate exactly which treatments are different.
- With more than two treatments, this creates a problem. Specifically, you must follow the ANOVA with additional tests, called post tests, to determine exactly which treatments are different and which are not.
- The Scheffe test and Tukey=s HSD are examples of post tests.
- These tests are done after an ANOVA where H_0 is rejected with more than two treatment conditions. The tests compare the treatments, two at a time, to test the significance of the mean differences.

Extensions to ANOVA

- There are also blocked ANOVA methodologies, etc. (analogous to the t-test blocking we looked at previously)
- However, we will not study them by hand but will examine them by computer in exercises.

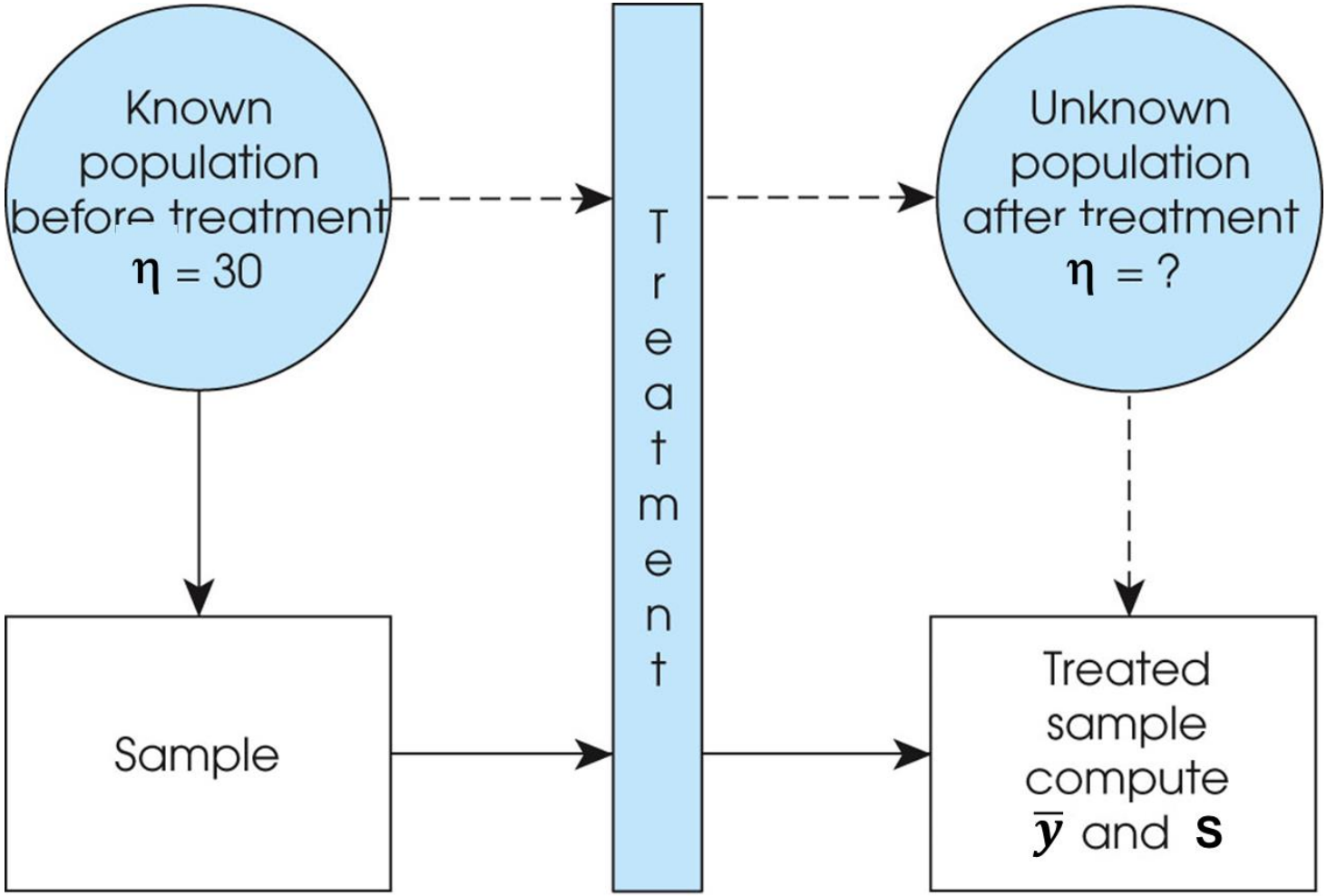
Pitfalls to consider with significance testing

- **Statistical significance does not guarantee experimental significance.**
- **Statistical significance does not imply a cause-effect relationship.**
- **Lack of statistical significance is not proof of the absence of an effect.**
- **Presence of statistical significance in one group and lack of statistical significance in another group ≠ a significant difference between the groups.**

Estimation

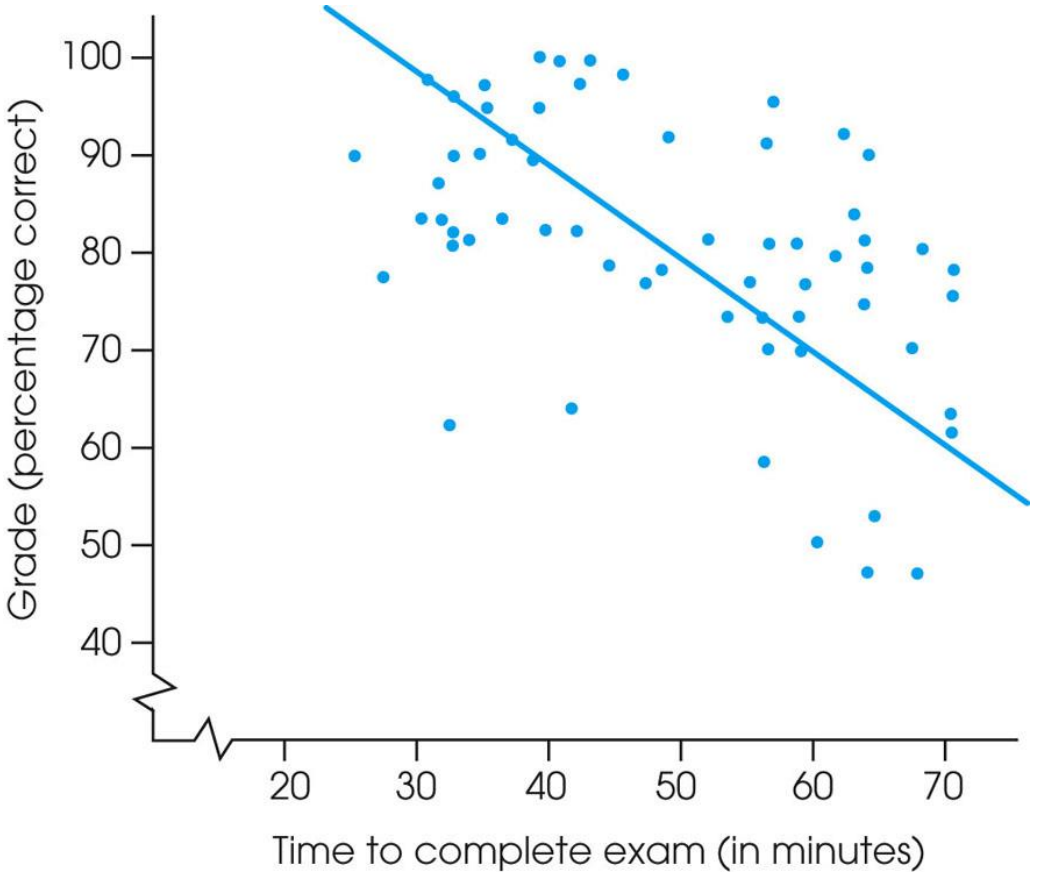
- In general terms, estimation uses a sample statistic as the basis for estimating the value of the corresponding population parameter.
- Although estimation and hypothesis testing are similar in many respects, they are complementary inferential processes.
- A hypothesis test is used to determine whether or not a treatment has an effect, while estimation is used to determine how much effect.
- This complementary nature is demonstrated when estimation is used after a hypothesis test that resulted in rejecting the null hypothesis.
- In this situation, the hypothesis test has established that a treatment effect exists and the next logical step is to determine how much effect.

Estimation


Known population before treatment
$\eta = 30$
Treatment
Unknown population after treatment
$\eta = ?$
Sample
Treated sample compute $\bar{y}$ and s

Correlations: Measuring and Describing Relationships

- **A correlation is a statistical method used to measure and describe the relationship between two variables.**
- **A relationship exists when changes in one variable tend to be accompanied by consistent and predictable changes in the other variable.**



Correlation

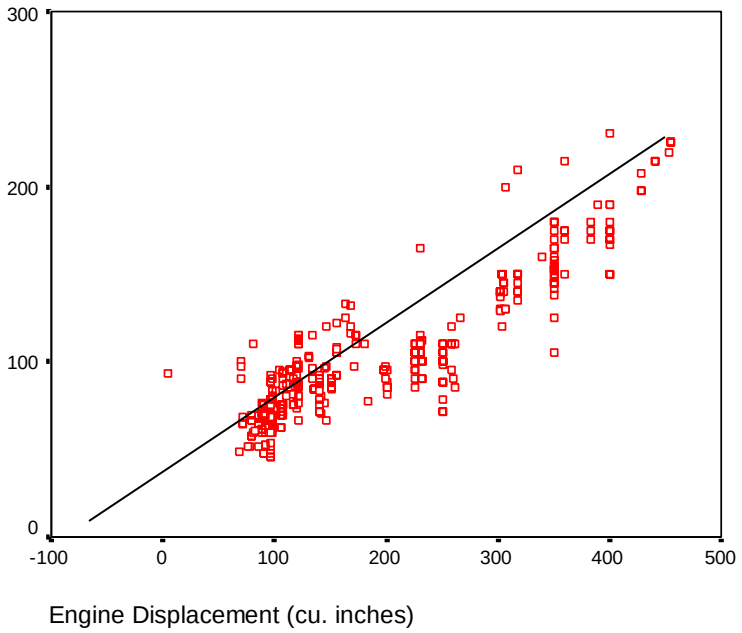
- A correlation typically evaluates three aspects of the relationship:
  - the direction
  - the form
  - the degree

- The direction of the relationship is measured by the sign of the correlation (+ or -). A positive correlation means that the two variables tend to change in the same direction; as one increases, the other also tends to increase. A negative correlation means that the two variables tend to change in opposite directions; as one increases, the other tends to decrease.

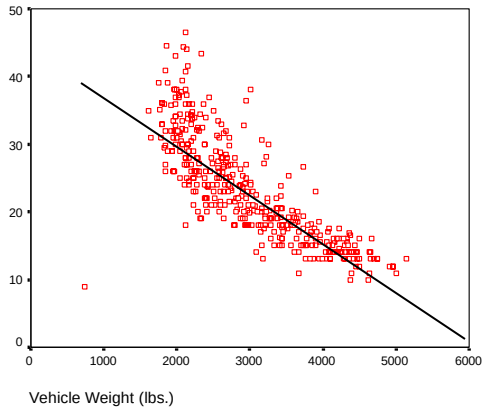
- The most common form of relationship is a straight line or linear relationship which is measured by the Pearson correlation.

- The degree of relationship (the strength or consistency of the relationship) is measured by the numerical value of the correlation. A value of 1.00 indicates a perfect relationship and a value of zero indicates no relationship.

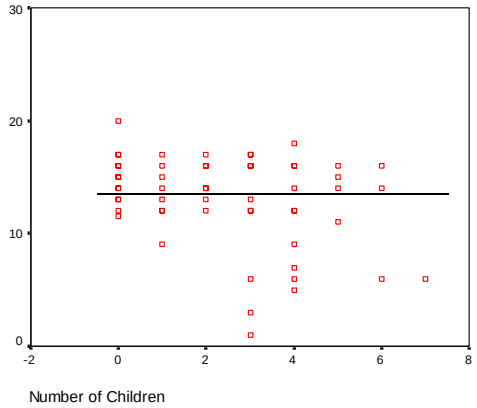
Various Types of Associations



Positive Relationship between X and Y

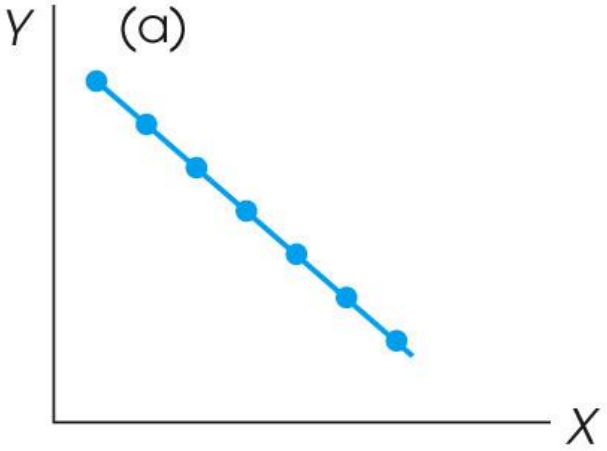


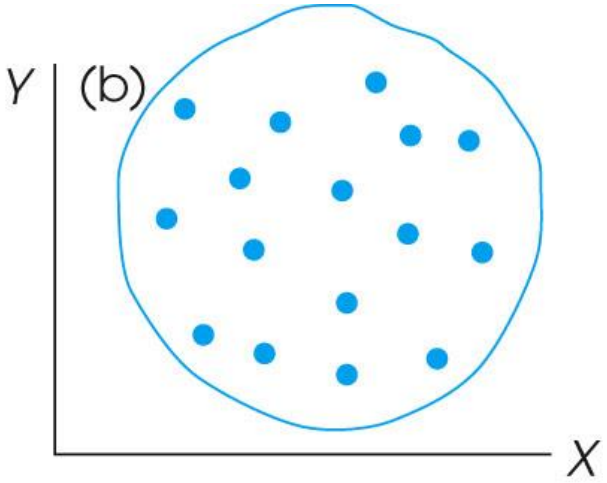
Strong negative Relationship between X and Y; points tightly clustered around line; nonlinear trend at lower weights

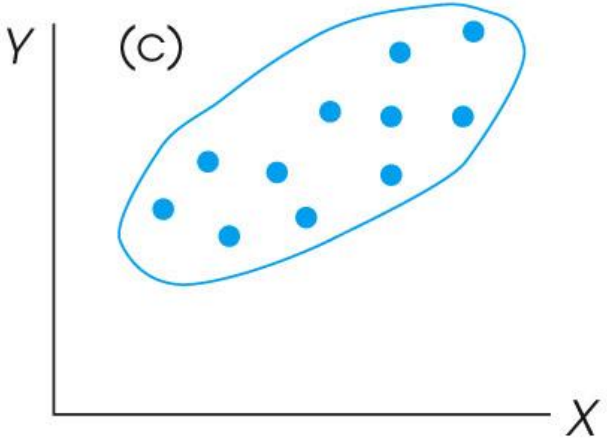


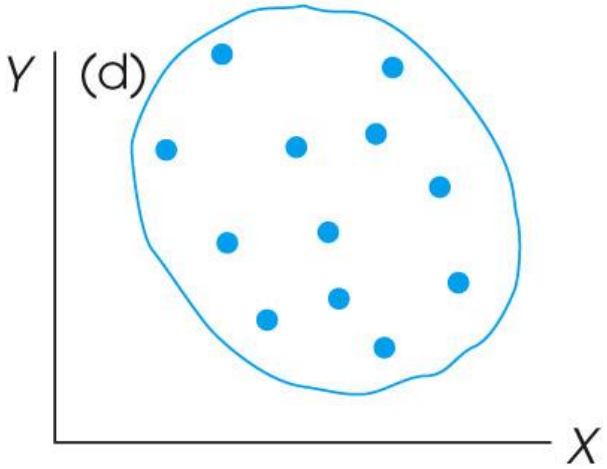
Essentially no relationship between X and Y; points loosely clustered around line

What degree of correlation to you expect for these data sets?









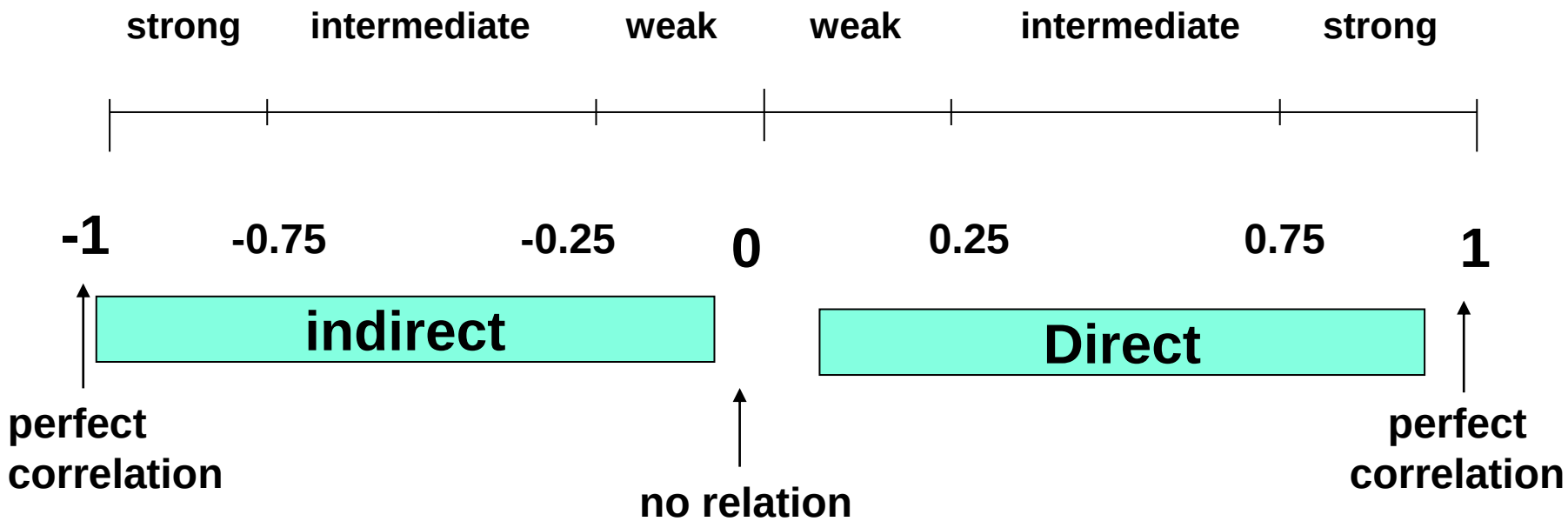
The Pearson Correlation

- **The Pearson correlation measures the direction and degree of linear (straight line) relationship between two variables.**
- **To compute the Pearson correlation, you first measure the variability of X and Y scores separately by computing S for the scores of each variable (S_X and S_Y).**
- **Then, the covariability (tendency for X and Y to vary together) is measured by the sum of products (SP).**
- **The Pearson correlation is found by computing the ratio,**

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{[\Sigma (X - \bar{X})^2][\Sigma (Y - \bar{Y})^2]}}$$

The Pearson coefficient

- The value of r ranges between (-1) and (+1)
- The value of r denotes the strength of the association as illustrated
  by the following diagram.



Related Measures of Association

- **The correlation coefficient is related to other types of measures of association:**
  - **The *partial correlation*, which measures the degree of association between two variables when the effects on them of a third variable is removed: what is the relationship between student achievement and dollars per student spent by the school district when the effect of parents' SES is removed**
  - **The *multiple correlation*, which measures the degree to which one variable is correlated with two or more other variables: how well can I predict student achievement knowing mean school district expenditure per pupil and parent SES**

An example: relation between shyness and speaking experience

The formula can be re-written as:

$$r = \frac{N \Sigma XY - \Sigma X \Sigma Y}{\sqrt{[N \Sigma X^2 - (\Sigma X)^2][N \Sigma Y^2 - (\Sigma Y)^2]}}$$

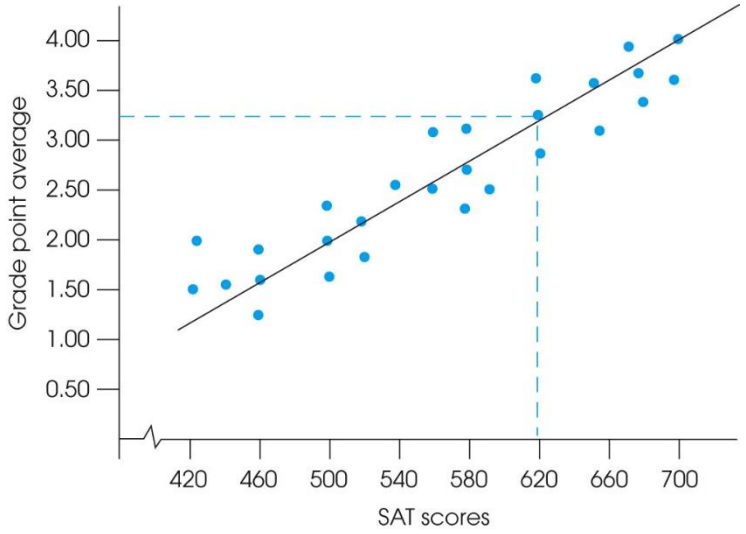
$$\frac{(6 \text{ X } 107) - 30\ (32)}{\sqrt{[6\ (230) - 30^2][6\ (226) - 32^2]}}$$

Shyness X	Speeches Y	XY	X^2	Y^2
0	8	0	0	64
2	10	20	4	100
3	4	12	9	16
6	6	36	36	36
9	1	9	81	1
10	3	30	100	9
30	32	107	230	226

r = -.797 (note crossproducts term in the numerator is negative) and R-square = .635

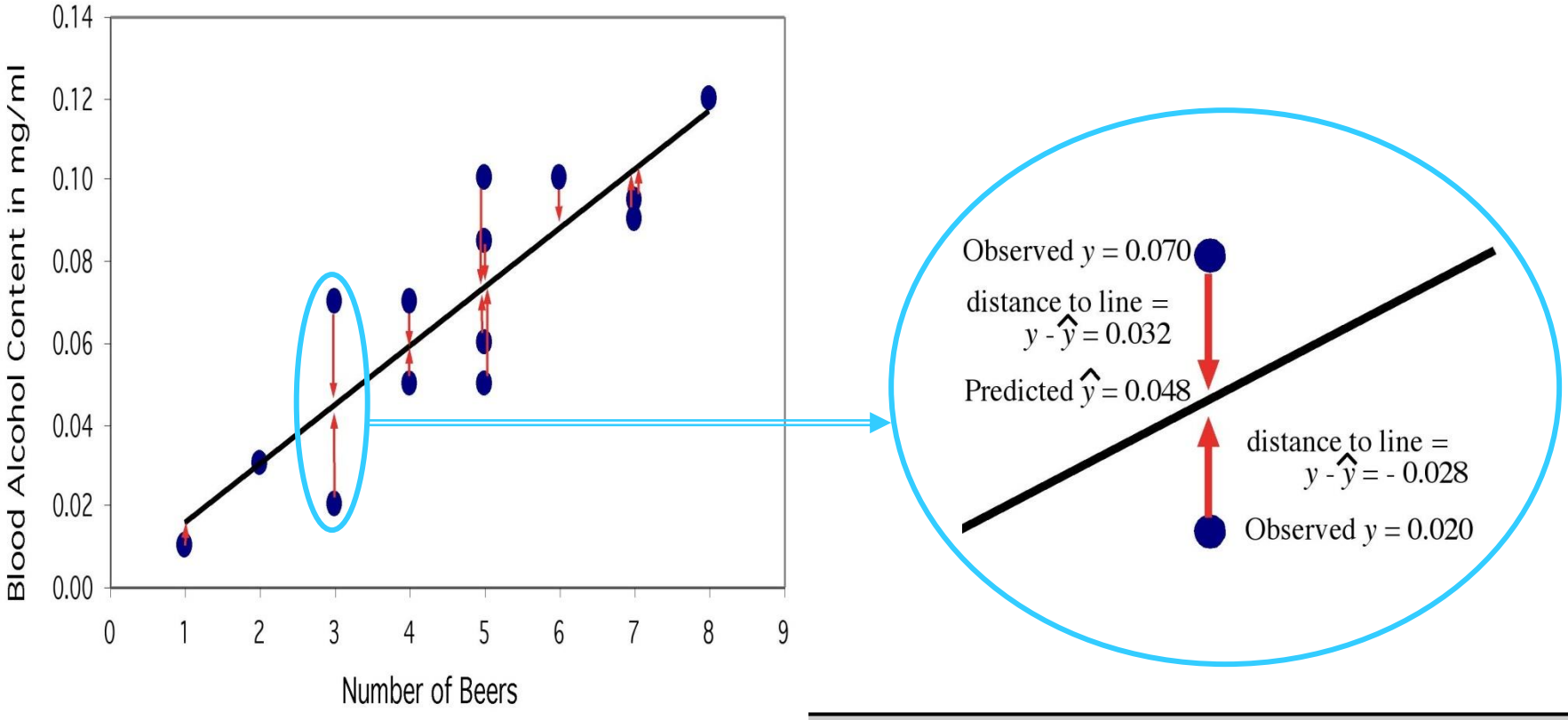
Introduction to Linear Regression

- The Pearson correlation measures the degree to which a set of data points form a straight line relationship.
- Regression is a statistical procedure that determines the equation for the straight line that best fits a specific set of data.
- How well a set of data points fits a straight line can be measured by calculating the distance between the data points and the line.
- The total error between the data points and the line is obtained by squaring each distance and then summing the squared values.
- The regression equation is designed to produce the minimum sum of squared errors.



The least-squares regression line

The least-squares regression line is the unique line such that the sum of the vertical distances between the data points and the line is zero, and the sum of the squared vertical distances is the smallest possible.



Finding the least-squares regression line

The slope of the regression line, *b*, equals:

$$b = r\frac{s_y}{s_x}$$

r **is the correlation coefficient between *x* and *y***
s_y is the standard deviation of the response variable *y*
s_x is the standard deviation of the explanatory variable *x*

The intercept, *a*, equals:

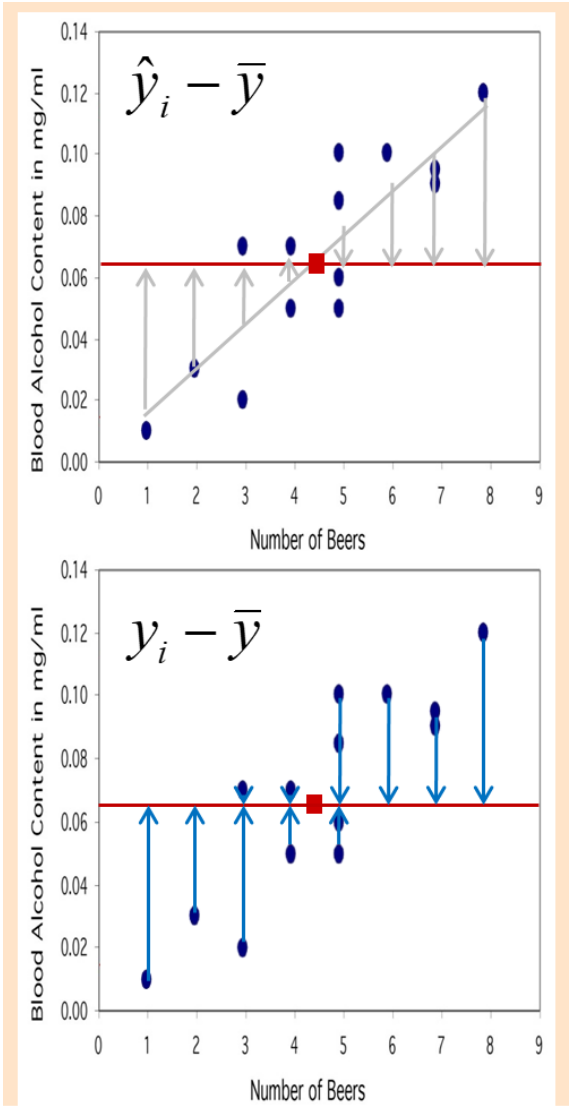
$$a = \bar{y} - b\bar{x}$$

$\bar{x}$ and $\bar{y}$ are the respective means of the *x* and *y* variables

The coefficient of determination, r^2

r^2, the coefficient of determination, is the square of the correlation coefficient.

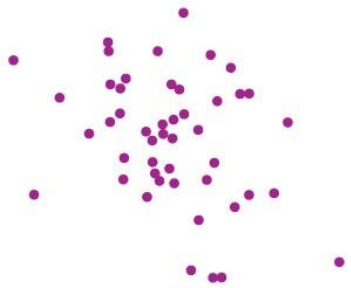
r^2 represents the fraction of the variance in y that can be explained by the regression model.



$r = 0.87$, so $r^2 = 0.76$

This model explains 76% of individual variations in BAC

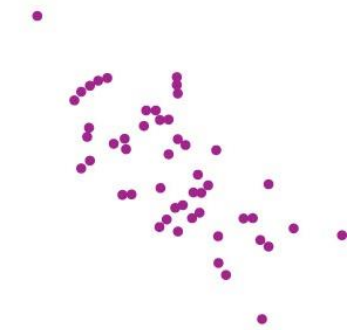
Considering what r means



Correlation $r = -0.3$

$r = -0.3$, $r^2 = 0.09$, or 9%

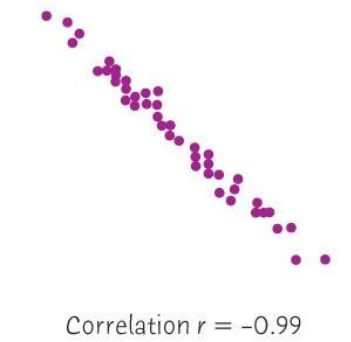
The regression model explains not even 10% of the variations in *y*.



Correlation $r = -0.7$

$r = -0.7$, $r^2 = 0.49$, or 49%

The regression model explains nearly half of the variations in *y*.



Correlation $r = -0.99$

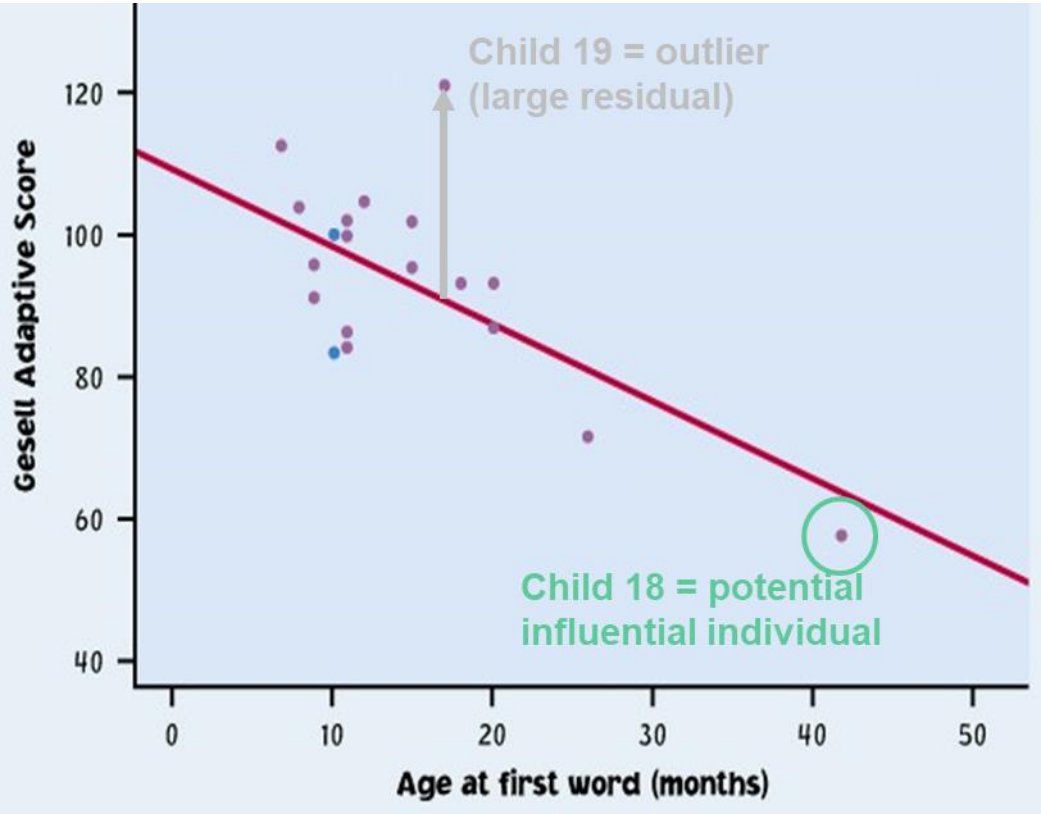
$r = -0.99$, $r^2 = 0.9801$, or ~98%

The regression model explains almost all of the variations in *y*.

Outliers and influential points

Outlier: An observation that lies outside the overall pattern.

"Influential individual": An observation that markedly changes the regression if removed. This is often an isolated point.



Child 19 is an outlier of the relationship (it is unusually far from the regression line, vertically).

Child 18 is isolated from the rest of the points, and might be an influential point.

Recognizing bad regression

1. Create scatterplot. Approximately linear?
2. Calculate r^2, the square of the correlation coefficient
3. Examine residual plot

Garbage In Garbage Out