Statistics - Part 2

Feb. 6, 2025

Statistics - Part 2

Feb. 6, 2025

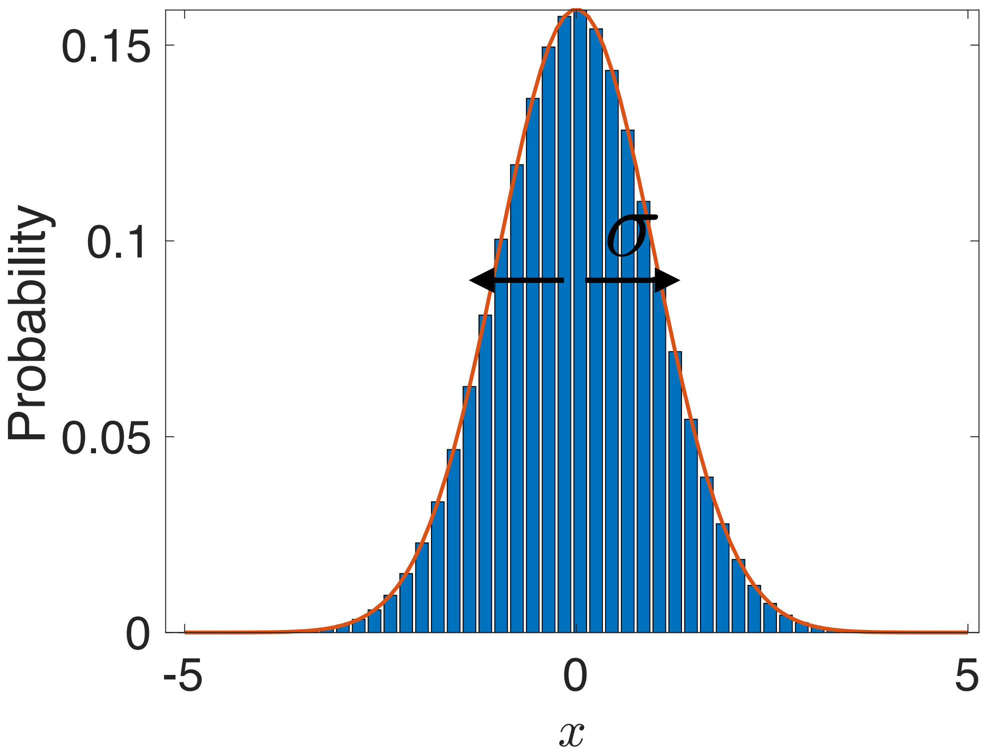
By the end of this lecture, you will be able to:

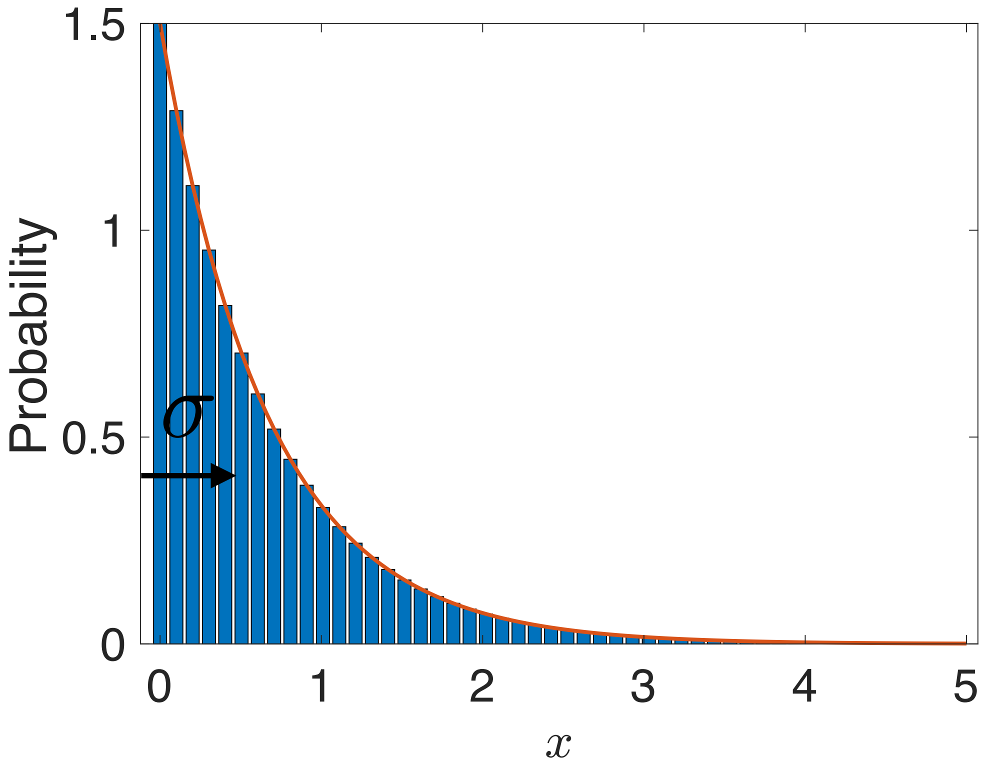
1. Give examples of statistical bias
2. Explain Simpson's paradox, Base rate fallacy, Will Rogers phenomenon and Berkson's paradox
3. Define null hypothesis and compute p-values

How do we compute confidence intervals?

Central limit theorem with more details

Let $\{x_1, x_2, \ldots, x_n\}$ be independent random samples drawn from a distribution of mean μ and standard deviation σ.





Recap of central limit theorem

Let $\{x_1, x_2, \ldots, x_n\}$ be independent random samples drawn from a distribution of mean μ and standard deviation σ. For large enough n (> 30), the distribution of the sample mean

$$\bar{x}_n = (x_1 + x_2 + \cdots + x_n)/n$$

is close to a normal distribution with mean μ and standard deviation $\sqrt{\sigma^2/n}$.

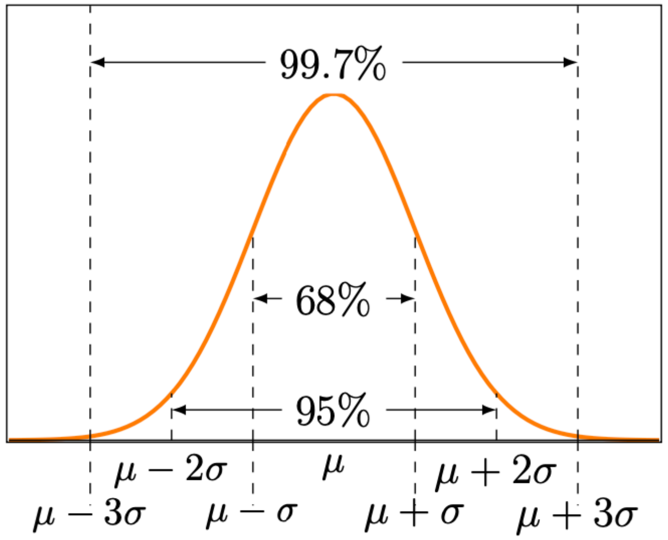
For smaller sample sizes, another distribution called the t-distribution must be used instead of the normal distribution

Normal Confidence Interval for Means

Example 2: In an orchard, there are hundreds of apples on the trees, and we want to measure the mean weight of the apples. You randomly choose just 46 apples and get the mean and the standard deviation of the weights as $\bar{x} = 86 \text{ g}, \ \sigma = 6.2 \text{ g}$ What is the 95% confidence interval?

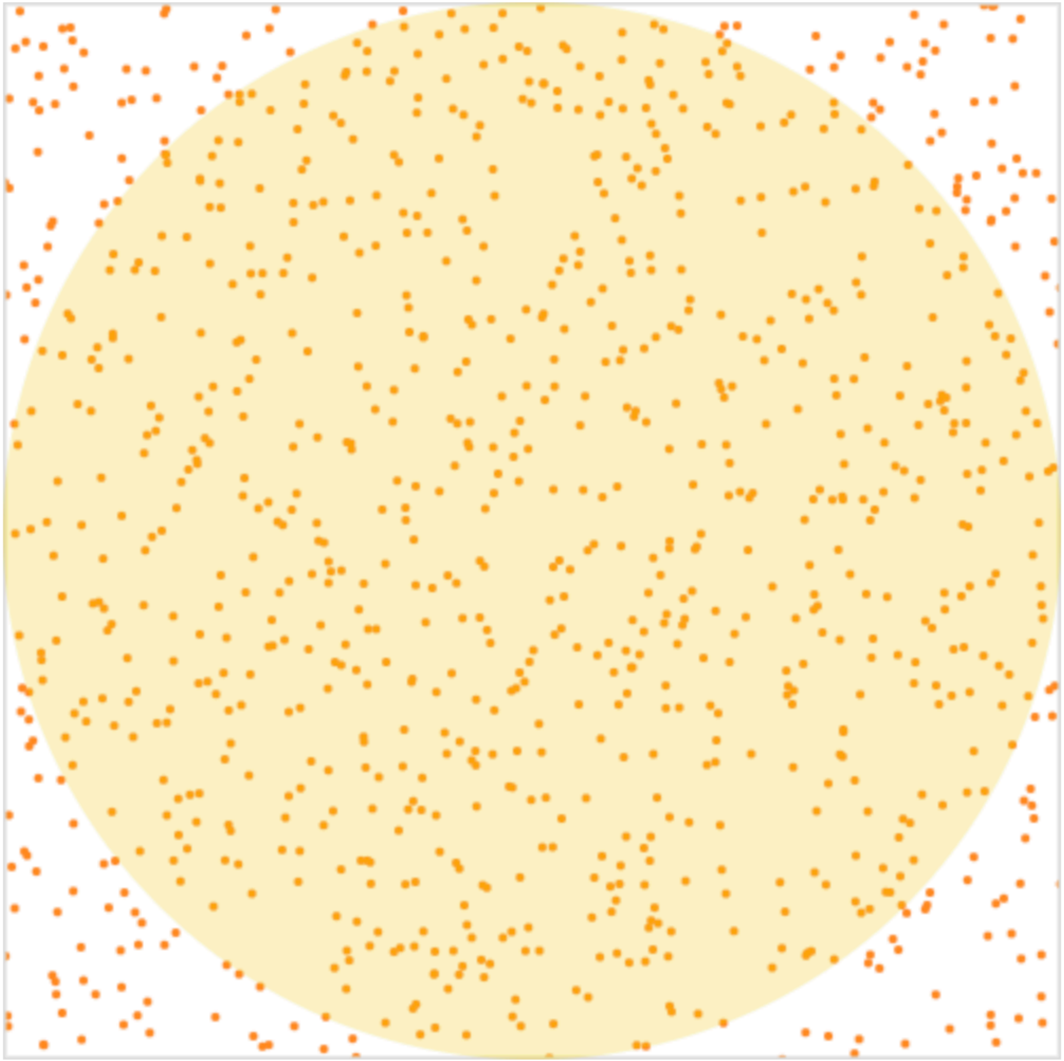
The standard deviation of the distribution of means $s = 6.2/\sqrt{46} = 0.9 \text{ g}$

3-sigma rule



For 95% confidence level, the margin of error = $2s = 1.8 \text{ g}$

The 95% confidence interval = (84.2, 87.8)



m: # of samples in the circle

n: # of samples dropped

m = 798, n = 1000

$$\hat{\pi} = \frac{4m}{n} = 3.192$$

Then, how do we calculate the confidence interval for our estimate of π ?

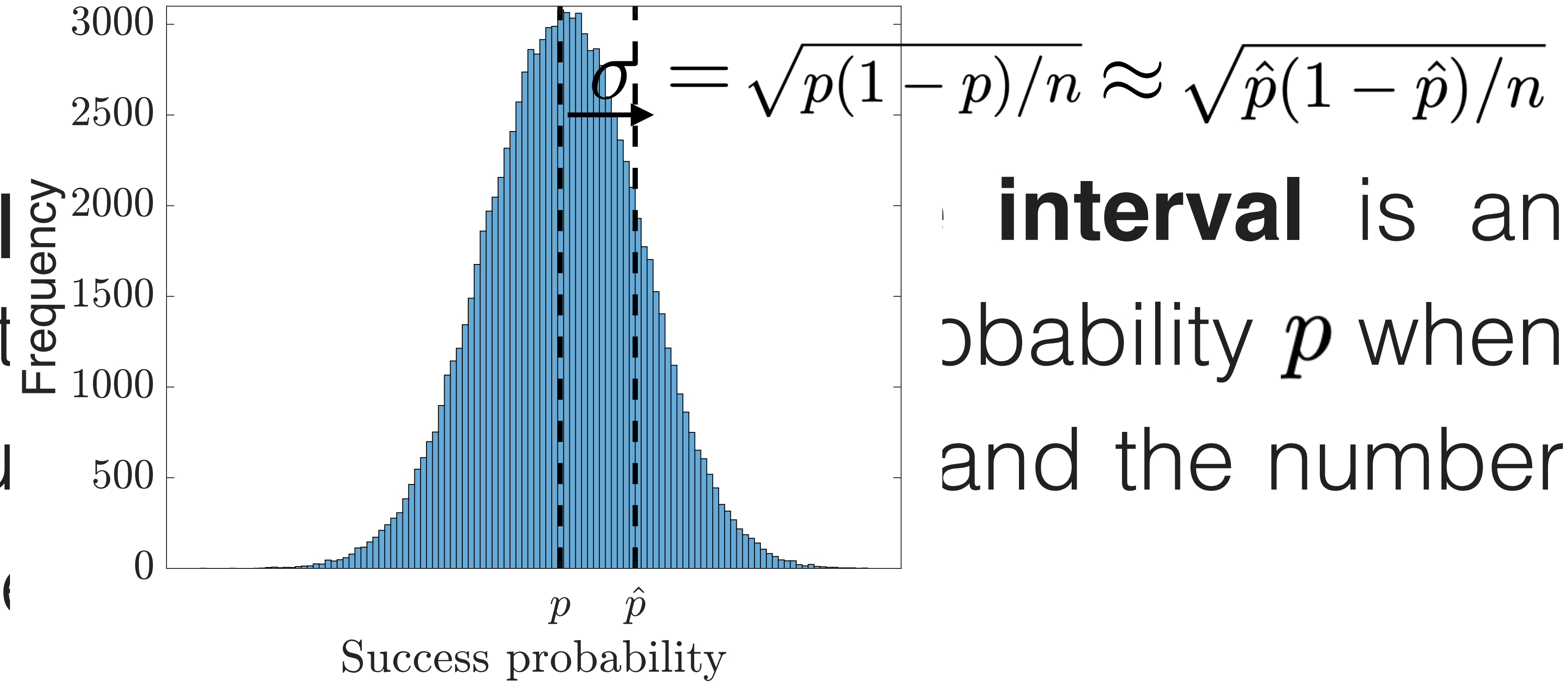
Binomial proportion confidence interval is an interval estimate of a success probability $\boldsymbol{p}$ when only the number of experiments $\boldsymbol{n}$ and the number of successes $\boldsymbol{n_s}$ are known

Let $\hat{p} = n_s/n$ be the estimate for p

If the sample size is not too small, the distribution of $\hat{p}$ is close to normal, with mean value p and standard deviation $\sqrt{p(1-p)/n}$

We can approximate this by $\sqrt{\hat{p}(1-\hat{p})/n}$
with the hope that p is not close to 0 or 1 and $\hat{p}$ is not too far from p

Binomial **interval** is an interval est obability p when only the nu and the number of successe



$\sigma = \sqrt{p(1-p)/n} \approx \sqrt{\hat{p}(1-\hat{p})/n}$

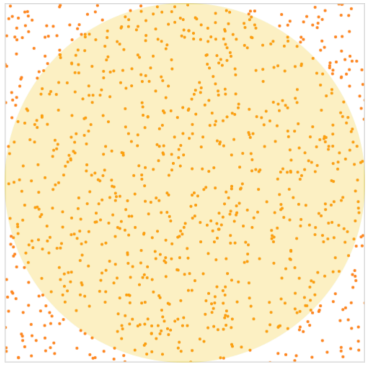
Let $\hat{p} = n_s/n$ be the estimate for p

If the sample size is not too small, the distribution of $\hat{p}$ is close to normal, with mean value p and standard deviation $\sqrt{p(1-p)/n}$

We can approximate this by $\sqrt{\hat{p}(1-\hat{p})/n}$ with the hope that p is not close to 0 or 1 and $\hat{p}$ is not too far from p

Example 3: Estimation for the value of π

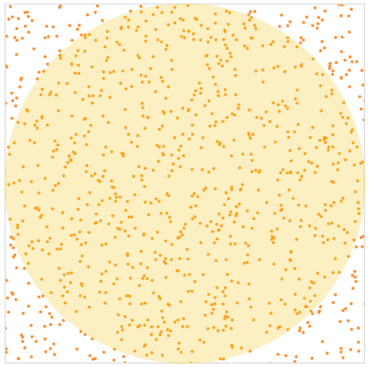
1. Compute the point estimate



$$\hat{p} = \frac{n_s}{n} = 0.798$$

n_s = 798, n = 1000

Example 3: Estimation for the value of π



n_s = 798, n = 1000

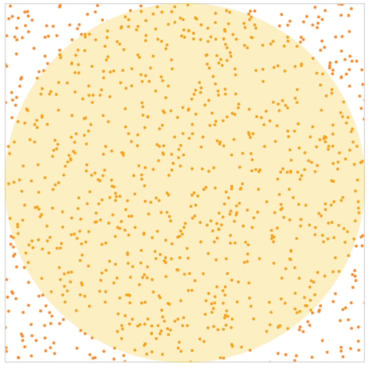
1. Compute the point estimate

$$\hat{p} = \frac{n_s}{n} = 0.798$$

2. Compute the standard deviation

$$s = \sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{\frac{0.798(1-0.798)}{1000}} = 0.013$$

Example 3: Estimation for the value of π



1. Compute the point estimate

$$\hat{p} = \frac{n_s}{n} = 0.798$$

n_s = 798, n = 1000

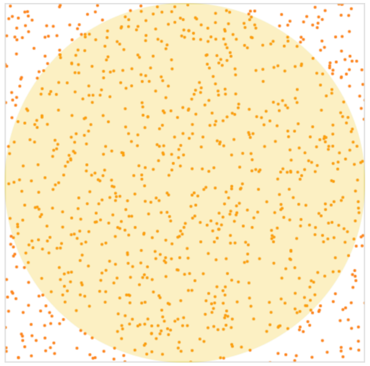
2. Compute the standard deviation

$$s = \sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{\frac{0.798(1-0.798)}{1000}} = 0.013$$

3.The 95% confidence interval for the success probability is

$$(\hat{p} - 2s, \hat{p} + 2s) = (0.772, 0.824)$$

Example 3: Estimation for the value of π



1. Compute the point estimate

$$\hat{p} = \frac{n_s}{n} = 0.798$$

n_s = 798, n = 1000

2. Compute the standard deviation

$$s = \sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{\frac{0.798(1-0.798)}{1000}} = 0.013$$

3.The 95% confidence interval for the success probability is

$$(\hat{p} - 2s, \hat{p} + 2s) = (0.772, 0.824)$$

Since $\pi = 4p$, the 95% confidence interval for π is (3.088, 3.296) (the conf. interval for p mult. by 4)

Takeaway

1. Means of Sample $\{x_1, x_2, \ldots, x_n\}$:

   Normal Confidence Interval

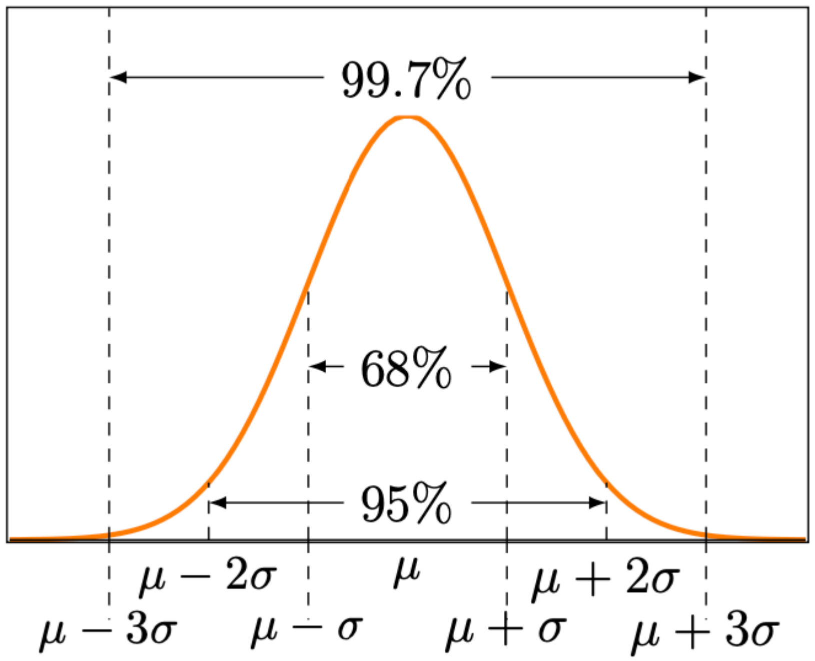
2. Success Rate p:

   Binomial Proportion Confidence Interval

Example 3’: You toss marbles into a cup. Out of 100 marbles, 52 land inside the cup. Compute a 99.7% confidence interval for the success probability.

100 marbles total, 52 land in the cup

3-sigma rule



1. Compute the point estimate

$$\widehat{p} = \frac{n_s}{n} = 0.52$$

2. Compute the standard deviation

$$s = \sqrt{\widehat{p}(1-\widehat{p})/n} = \sqrt{\frac{0.52(1-0.52)}{100}} = 0.05$$

3.The 99.7% confidence interval for success probability is

$$(\widehat{p} - 3s, \widehat{p} + 3s) = (0.37, 0.67)$$

Statistical Bias

Common sampling bias

When a sample does not have the same characteristics as the population, we say this is a biased sample.

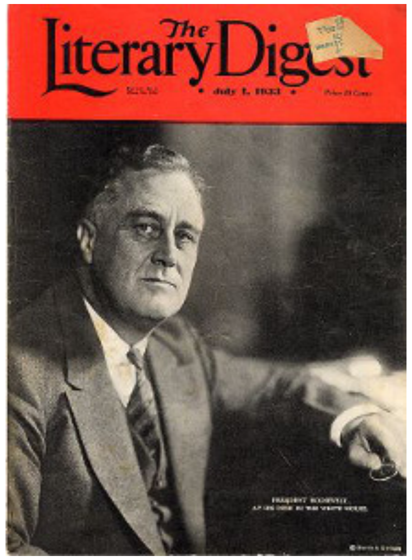
Common sampling bias

When a sample does not have the same characteristics as the population, we say this is a biased sample.

Example: poll by the magazine “The Literary Digest” for the 1936 presidential election.

Prediction: Alf Landon beats Franklin D Roosevelt by 57% to 43%.

Result: Roosevelt beats Landon by 62% to 38%



What went wrong in the 1936 poll?

Prediction: Alf Landon beats Franklin D Roosevelt by 57% to 43%.
Result: Roosevelt beats Landon by 62% to 38%

Questionnaires were sent out using lists of phone numbers, drivers' registrations, and country club memberships. 24% of those polled responded.

What went wrong?

What went wrong in the 1936 poll?

Questionnaires were sent out using lists of phone numbers, drivers' registrations, and country club memberships. 24% of those polled responded.

- The people polled were wealthy and the election was at the height of the Great Depression
- A large fraction of polled did not respond. Typically, those with strong feelings respond
- …

What went wrong in the 1936 poll?

Questionnaires were sent out using lists of phone numbers, drivers' registrations, and country club memberships. 24% of those polled responded.

- The people polled were wealthy and the election was at the height of the Great Depression
- A large fraction of polled did not respond. Typically, those with strong feelings respond
- …

Note that the poll was enormous (10 million questionnaires!), but bias still made the result useless!

Common sampling bias

When a sample does not have the same characteristics as the population, we say this is a biased sample.

We will talk about four kinds of bias:

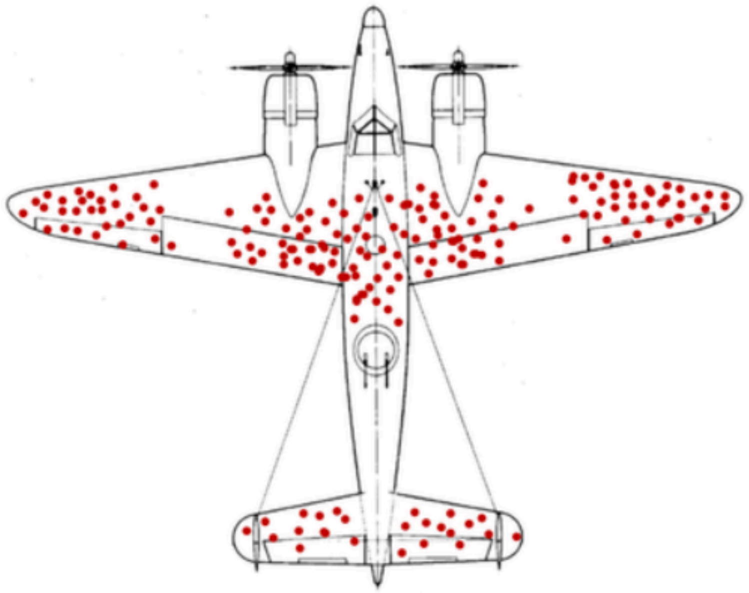
1. Survival bias
2. Self-selection bias
3. Confirmation bias
4. Undercoverage bias

Common sampling bias

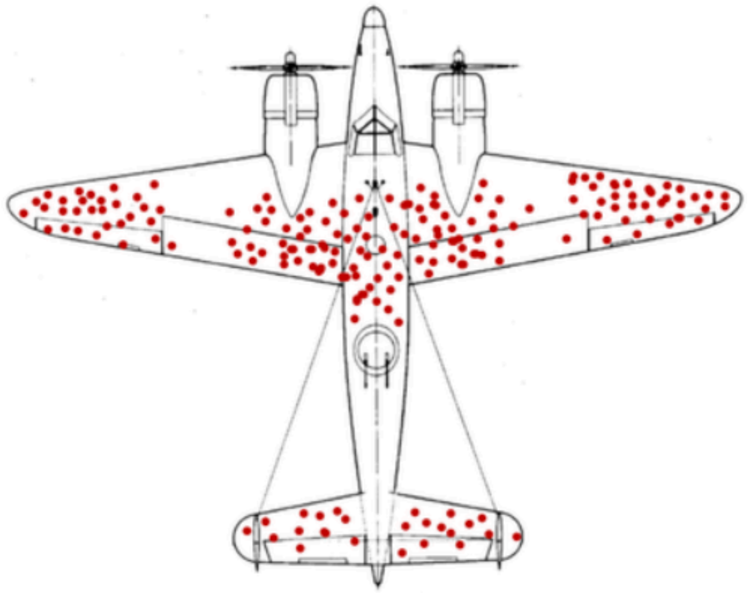
When a sample does not have the same characteristics as the population, we say this is a biased sample.

Survival bias: only the portion of the population that has survived some process can be sampled

Example 1: In scientific journals, there is strong publication bias towards positive results. Successful research outcomes are published far more often than null findings.



Example 2: During World War II, the statistician Abraham Wald examined the damage done to aircrafts that had returned from missions. The US military previously concluded that the most-hit areas of the plane needed additional armor.



Example 2: During World War II, the statistician Abraham Wald examined the damage done to aircrafts that had returned from missions. The US military previously concluded that the most-hit areas of the plane needed additional armor. Wald instead recommended adding armor to the areas that showed the least damage.

Self-selection bias: People with specific characteristics are more likely to agree to take part in a study than others. This often leads to a polarization of responses with extreme perspectives being given a disproportionate weight in the summary.

Example: people who have strong opinions or substantial knowledge may be more willing to spend time answering a survey than those who do not.

Confirmation bias: People display this bias when they select information that supports their views.

Example 1: In social media, personalized search displays to individuals only information they are likely to agree with, while excluding opposing views.

Example 2: The decision made by a doctor may be strongly influenced by the disorders described in a recently-read paper, without considering multiple possibilities based on evidence.

Undercoverage bias: Some members of a population are inadequately represented in the sample.

Example 1: Administering general national surveys online may miss groups with limited internet access, such as the elderly and lower-income households.

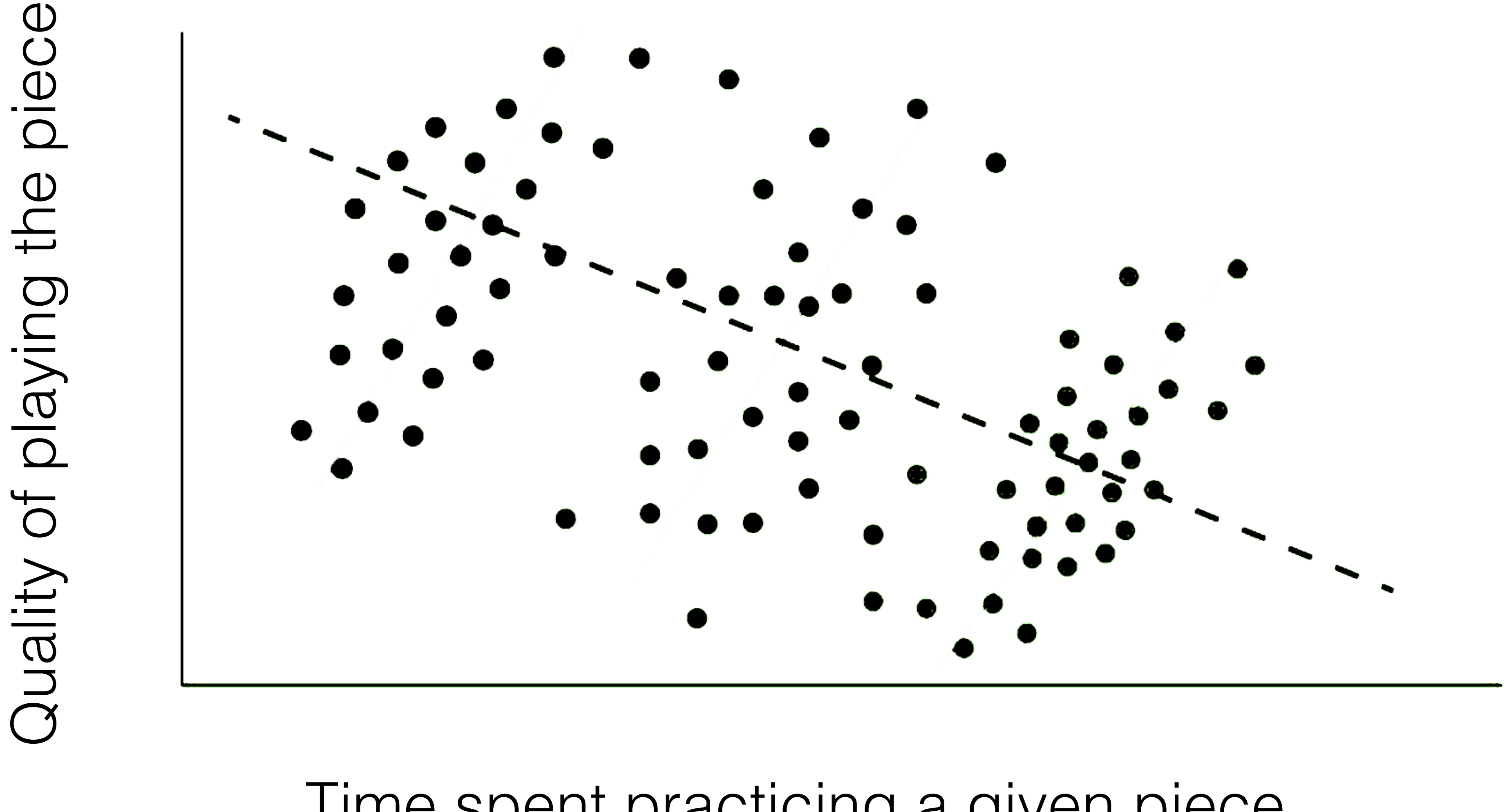
Example 2: Researchers want to know what citizens in a particular city think of a new traffic law so they give out a questionnaire to people that walk by at a local mall.

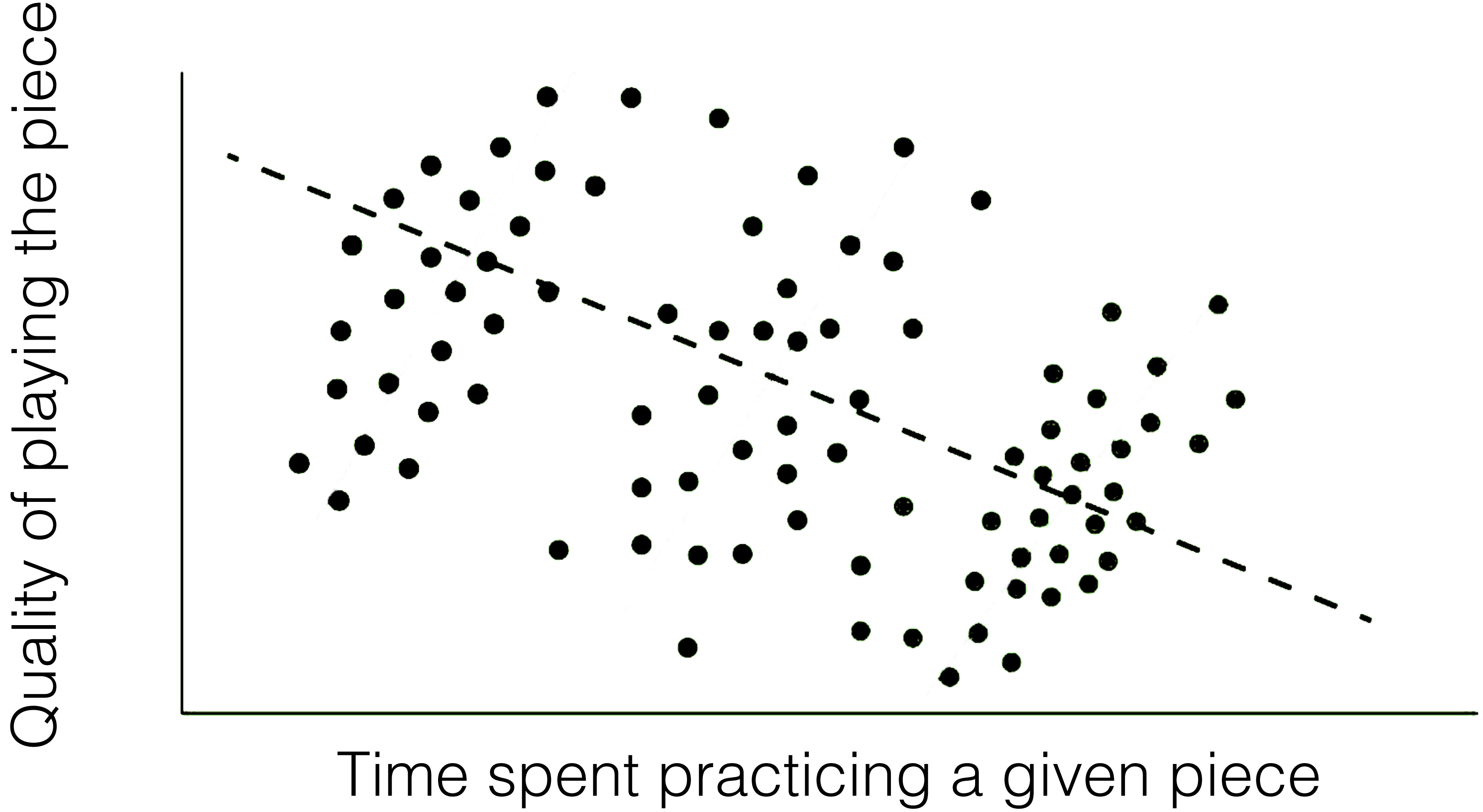
Statistical Paradoxes

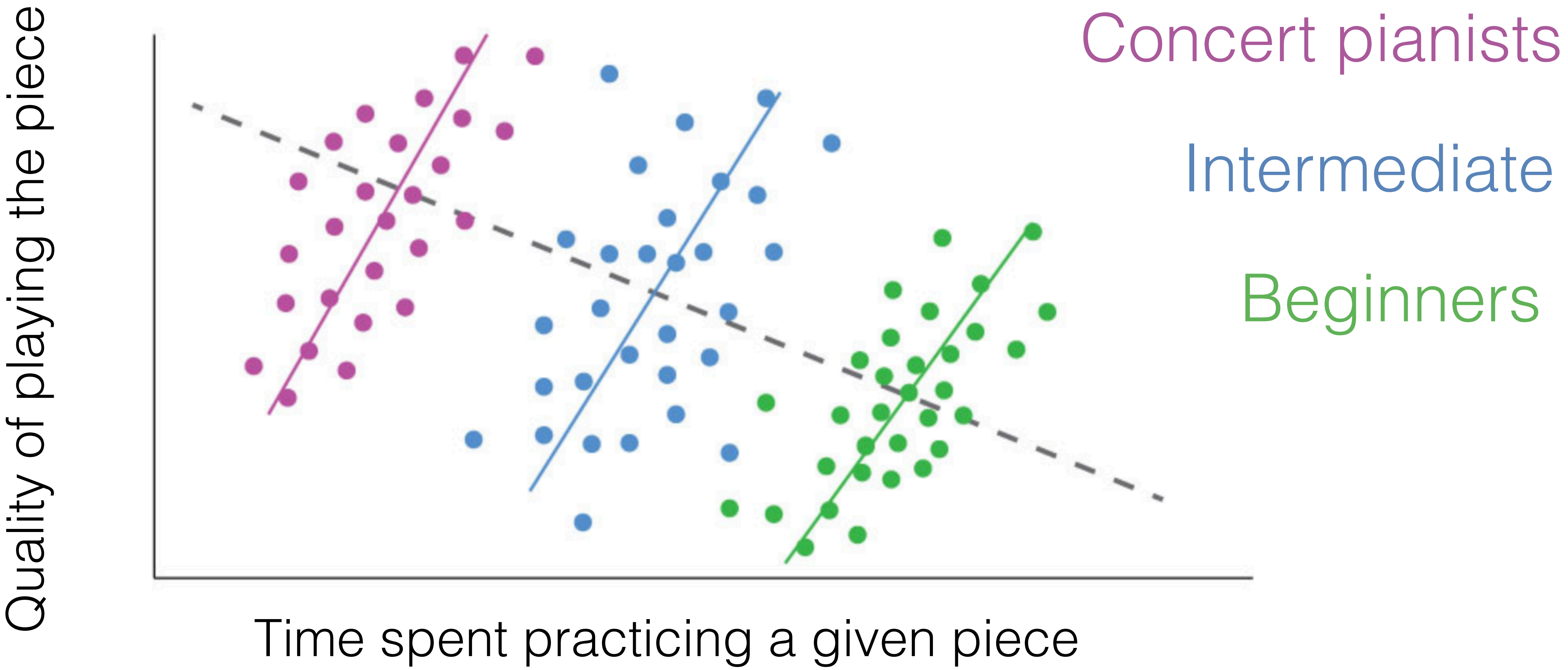
“There are three kinds of lies:
lies, damned lies, and statistics.”
-Mark Twain, 1907

Simpson's paradox

Does practice make perfect for musical instruments?

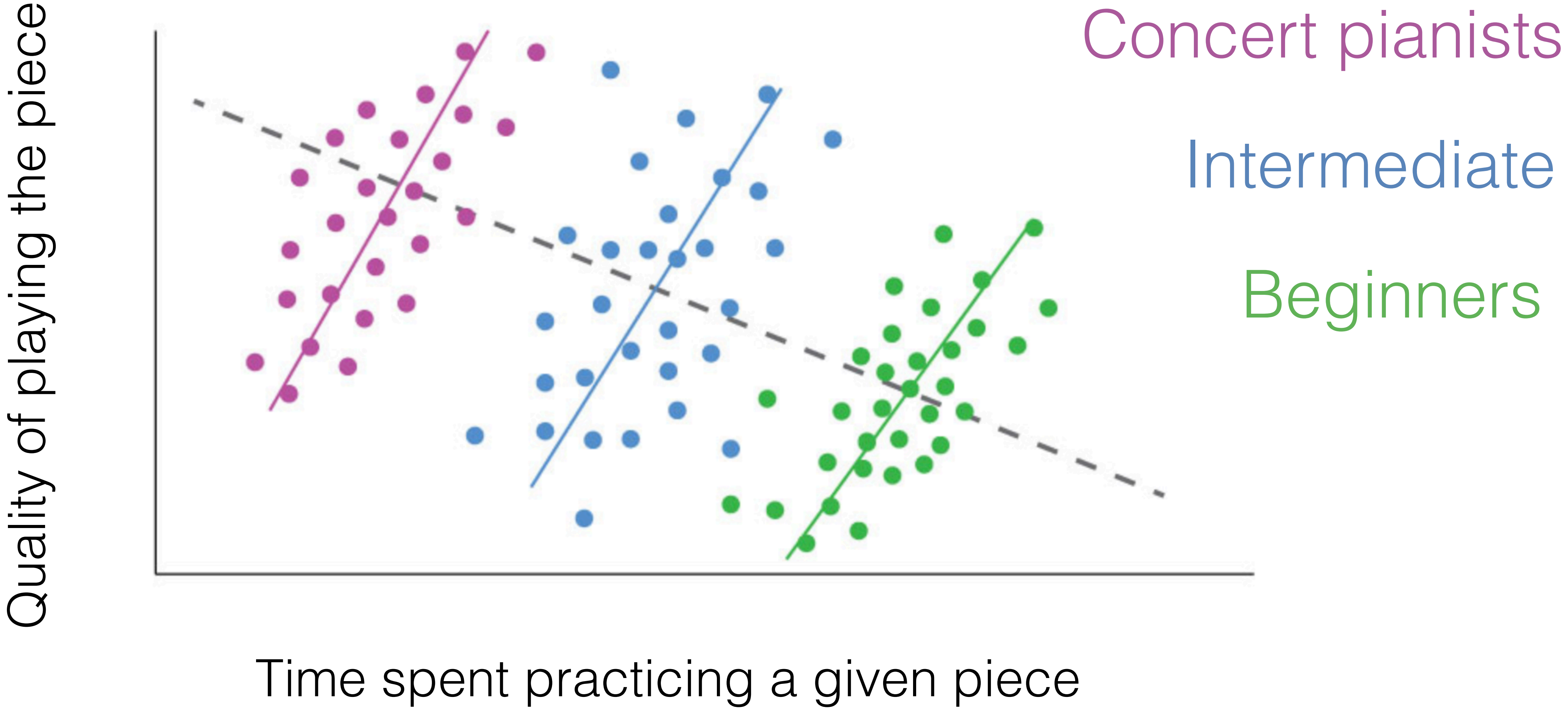



Quality of playing the piece
Time spent practicing a given piece


Quality of playing the piece
Time spent practicing a given piece
Concert pianists
Intermediate
Beginners

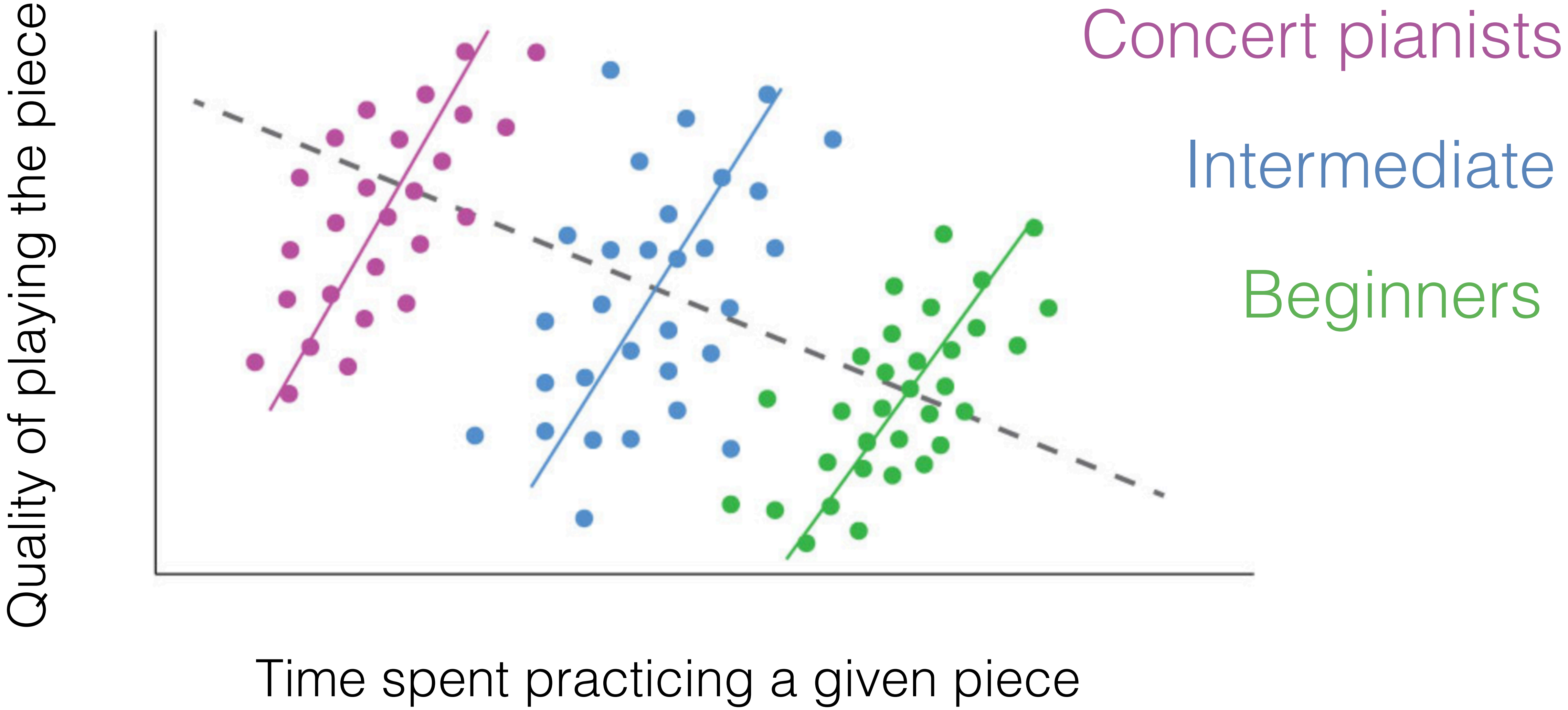
Simpson’s paradox

A trend appears in several different groups of data but disappears or reverses when these groups are combined.

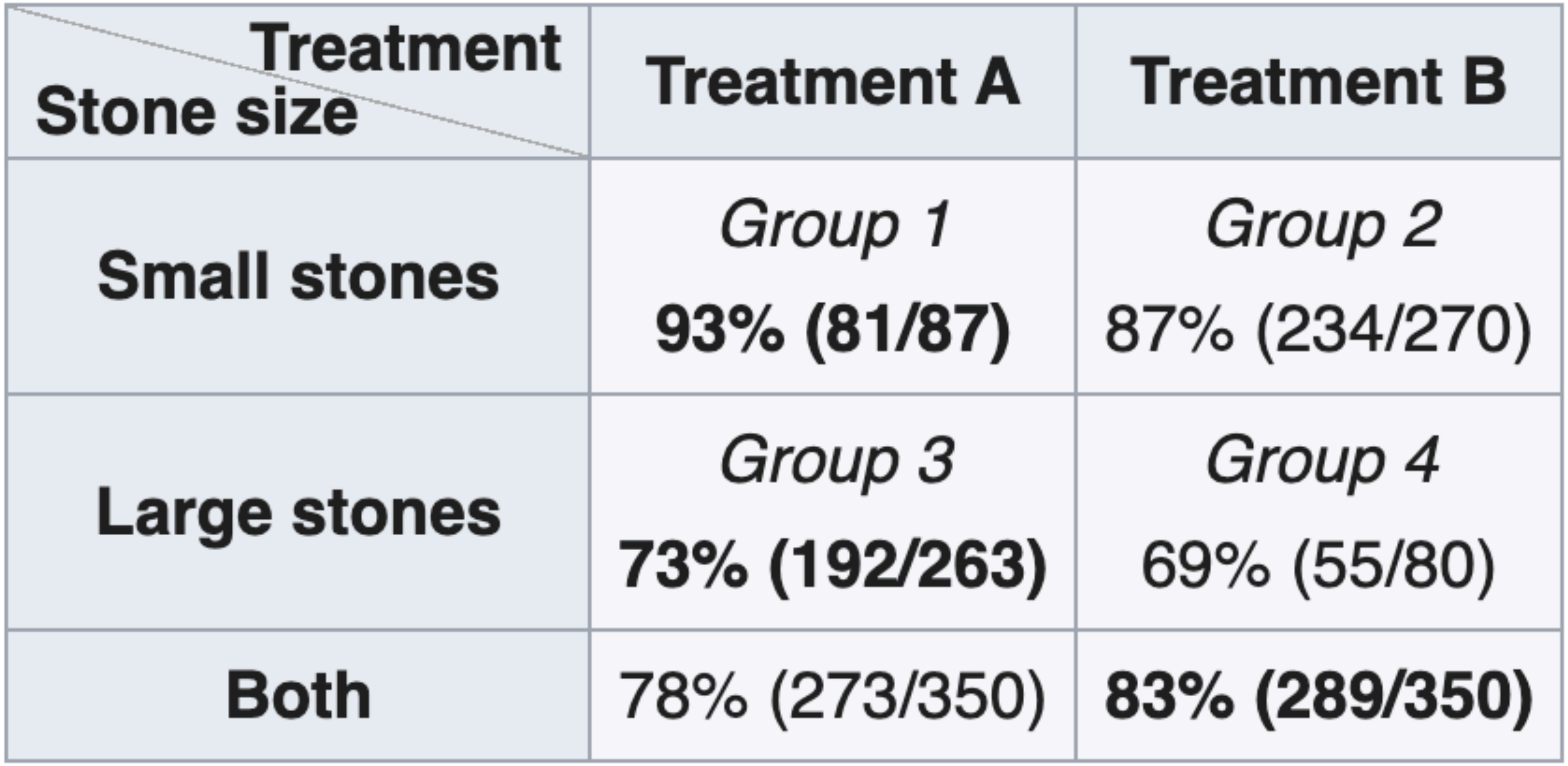


Simpson’s paradox

The skill level is a “hidden variable” that reverses the conclusion unless it is taken into account!

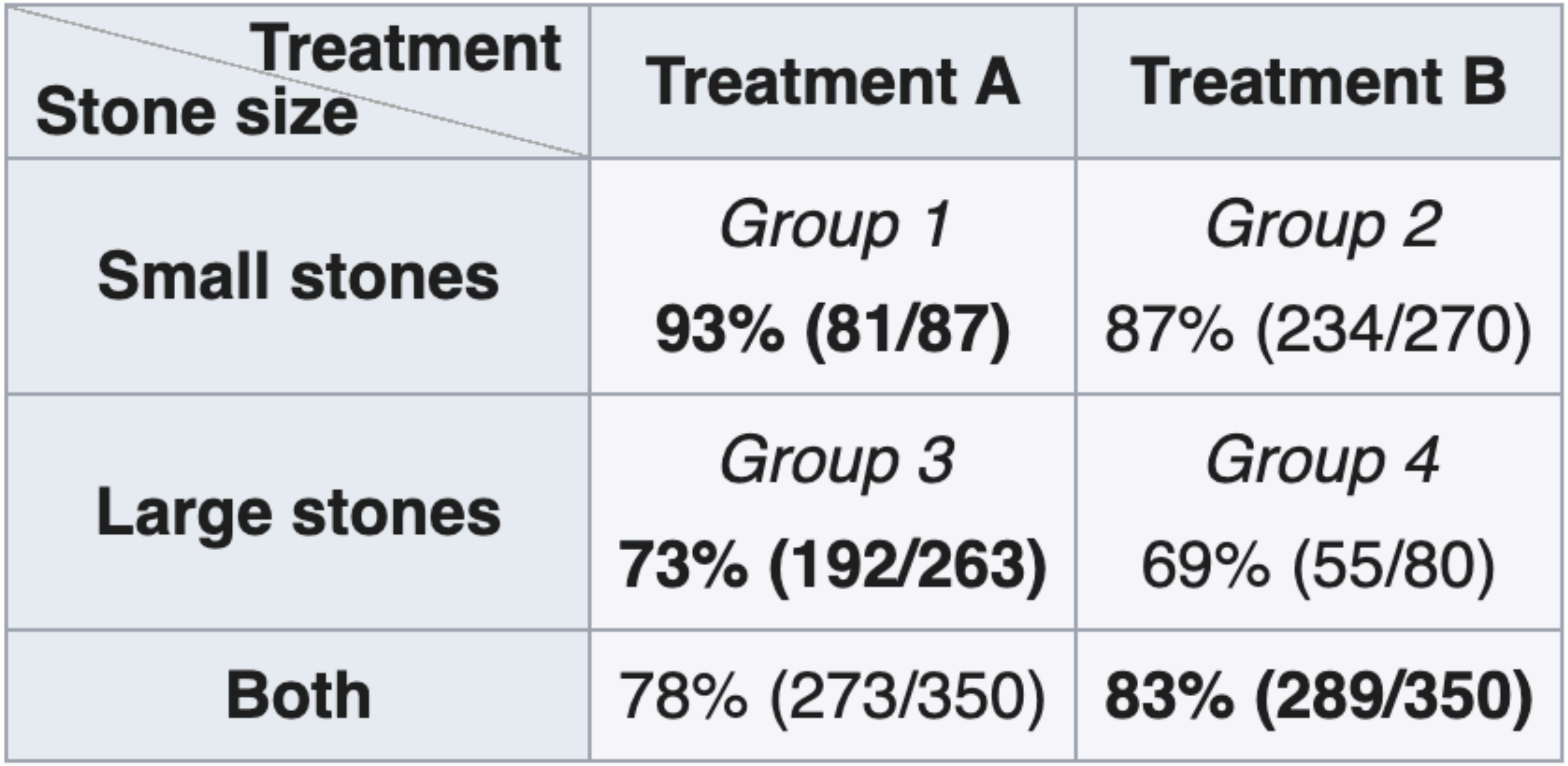


Example 1: Kidney stone treatment



Treatment Stone size	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

The kidney stone size is a hidden variable!



Treatment Stone size	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

Simpson's paradox arises because:

1. The severity of the the condition (small or large stone) has stronger impact on the outcome than the treatment
2. The groups are lopsided in size

Example 2: University admission

	Total		
	admit	denied	
Males	91	19	83%
Females	19	91	17%

Lawsuit against UC Berkeley in 1973

Example 2: University admission

	Math			Chemistry			Total		
	admit	denied		admit	denied		admit	denied	
Males	90	10	90%	1	9	10%	91	19	83%
Females	9	1	90%	10	90	10%	19	91	17%

Major is a “hidden variable”!

Example 2: University admission

	Math			Chemistry			Total		
	admit	denied		admit	denied		admit	denied	
Males	90	10	90%	1	9	10%	91	19	83%
Females	9	1	90%	10	90	10%	19	91	17%

Major is a “hidden variable”!

Example 2: University admission

	Math			Chemistry			Total		
	admit	denied		admit	denied		admit	denied	
Males	90	10	90%	1	9	10%	91	19	83%
Females	9	1	90%	10	90	10%	19	91	17%

Appearance of bias where there is none since more females applied to the harder major

	Math			Chemistry			Total		
	admit	denied		admit	denied		admit	denied	
Males	20	5	80%	40	55	42%	60	60	50%
Females	50	20	71%	10	40	20%	60	60	50%

In this example, Simpson’s paradox has masked the existence of bias

Base Rate Fallacy

If presented with related base rate information (i.e., general information on prevalence) and specific information (i.e., information pertaining only to a specific case), people tend to ignore the base rate in favor of the individuating information, rather than correctly integrating the two.

Example 1: When we hear that someone loves music, we might think it's more likely that the person is a professional musician than an accountant.

Base rate fallacy

Example 1: When we hear that someone loves music, we might think it's more likely that the person is a professional musician than an accountant.

Makes sense to assume:

P(loves music | musician) > P(loves music | accountant)

But not enough information to determine

P(musician | loves music) vs P(accountant | loves music)

It depends on how many accountants there are vs musicians

Example 2:

In a city, 85% of cabs are blue and the rest are green. One night, a cab is involved in a hit and run accident. A witness claims the cab was green, however later tests show that they only correctly identify the color of the cab at night 80% of the time. When asked what the probability is that the cab involved in the hit and run was green, people tend to answer that it is 80%.

Example 2:

In a city, **100%** of cabs are blue and the **0%** are green. One night, a cab is involved in a hit and run accident. A witness claims the cab was green, however later tests show that they only correctly identify the color of the cab at night 80% of the time. When asked what the probability is that the cab involved in the hit and run was green, people tend to answer that it is 80%.

Example 2:

In a city, **0%** of cabs are blue and the **100%** are green. One night, a cab is involved in a hit and run accident. A witness claims the cab was green, however later tests show that they only correctly identify the color of the cab at night 80% of the time. When asked what the probability is that the cab involved in the hit and run was green, people tend to answer that it is 80%.

Example 2:

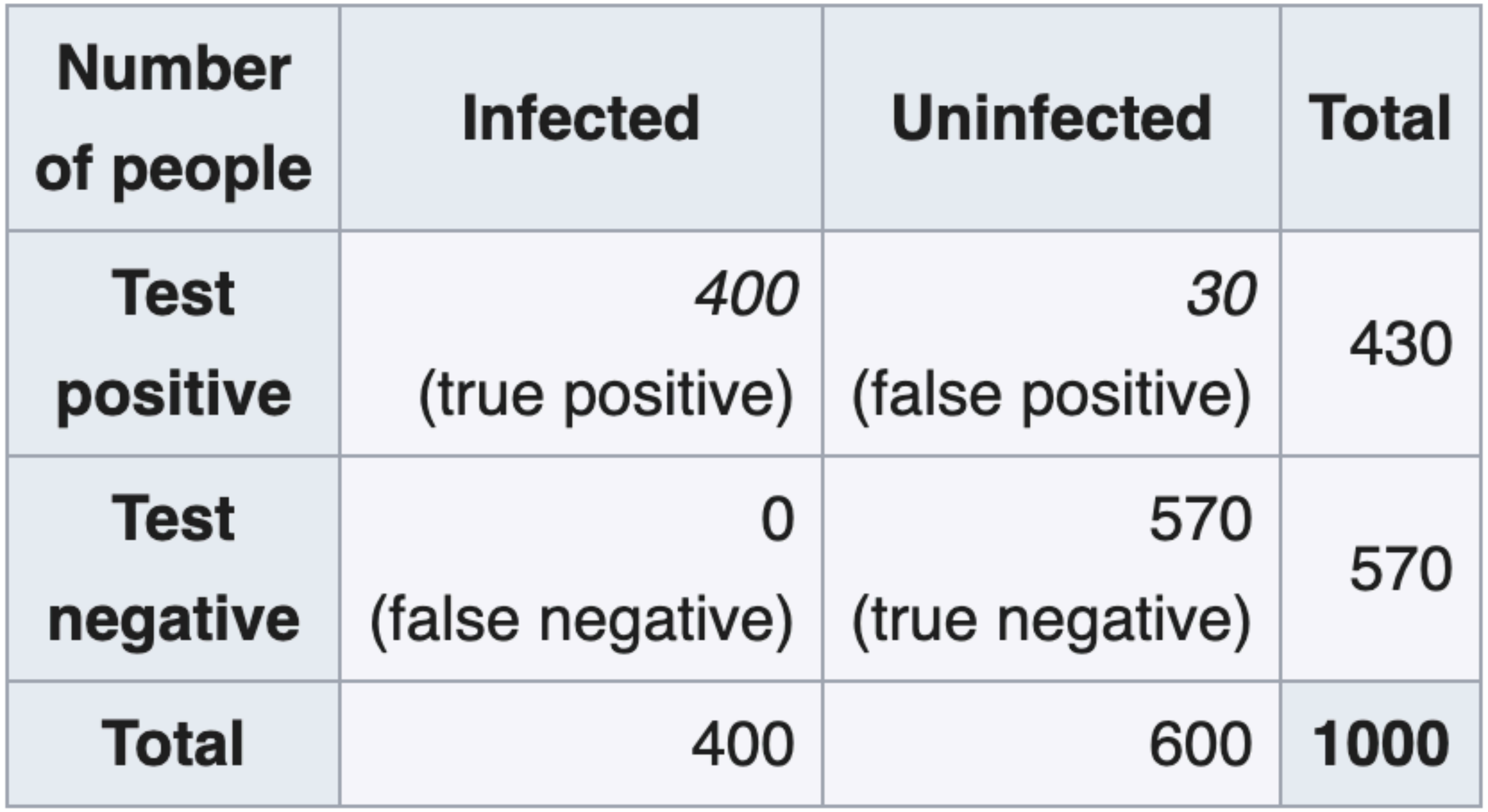
Compute P(it is green | identified green) using Bayes' rule!

Know P(green) = 0.15
P(identified green | it is green) = 0.8
P(identified green | it is blue) = 0.2

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Example 3:

A test on an infectious disease has 100% sensitivity and 95% specificity. The expected outcome of the 1,000 tests on population *A*, which has an infection rate of 40%, would be

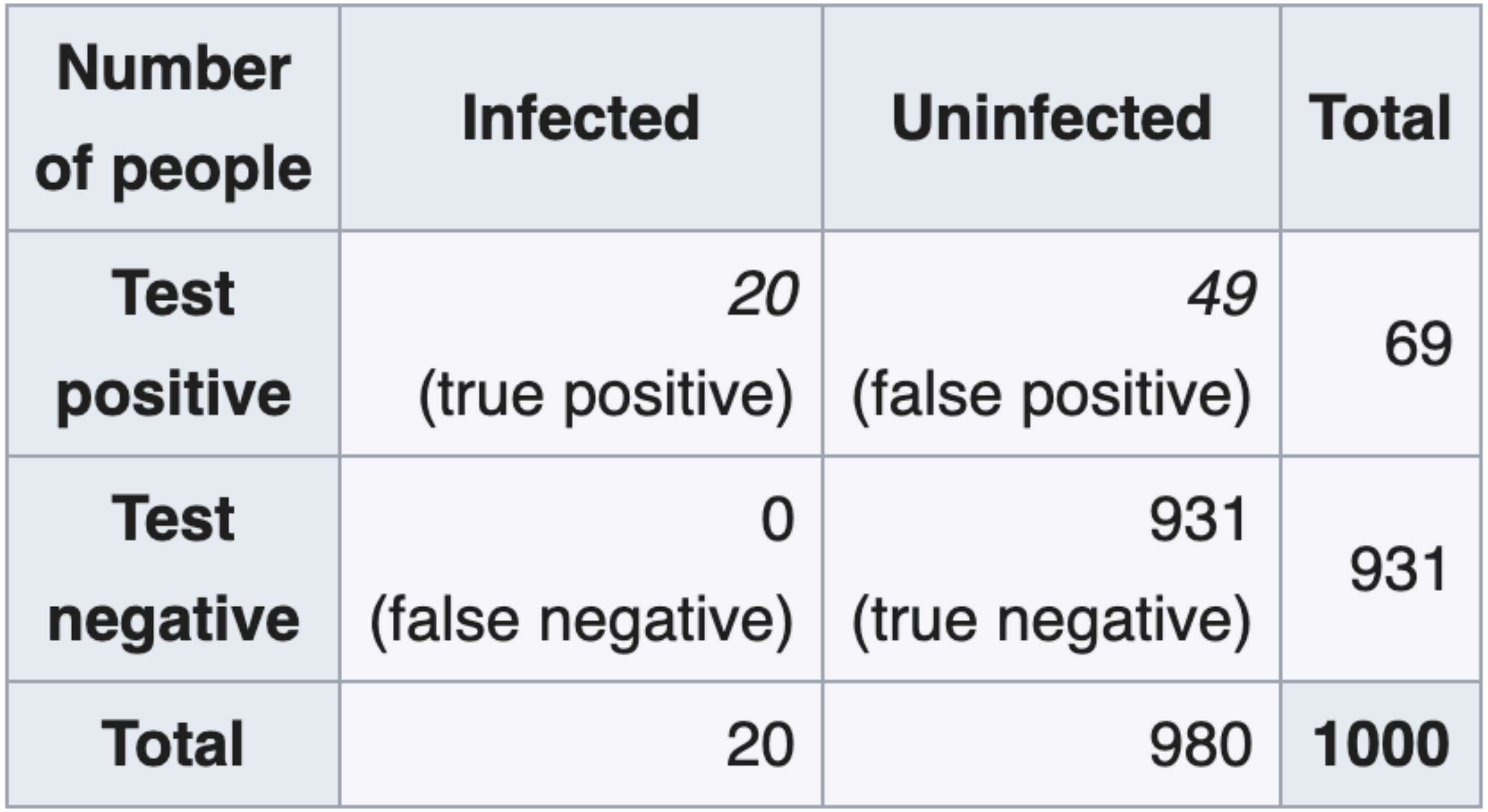


Number of people	Infected	Uninfected	Total
Test positive	*400* (true positive)	*30* (false positive)	430
Test negative	0 (false negative)	570 (true negative)	570
Total	400	600	**1000**

More true positives than false positives

Sensitivity = probability that test shows positive given that you are infected
Specificity = probability that test shows negative given that you are not infected

Now consider the same test applied to population B, in which only 2% is infected. The expected outcome of 1000 tests on population B would be:



Number of people	Infected	Uninfected	Total
Test positive	*20* (true positive)	*49* (false positive)	69
Test negative	0 (false negative)	931 (true negative)	931
Total	20	980	**1000**

More false positives than true positives

A tester with experience of group A might find it a paradox that in group B, a result that had usually correctly indicated infection is now usually a false positive.

Hypothesis testing

Hypothesis testing

Consider the following example:

48 bank supervisors (all male) were given the same personnel file and asked whether the person should be promoted or not. The files are identical, except that 24 of the files were assigned to belong to male employees and 24 to females. Of the 48 files, 35 were promoted, 21 of which belonged to males and the rest belonged to females.

The percentage of men promoted $= \dfrac{21}{24} = 88\%$

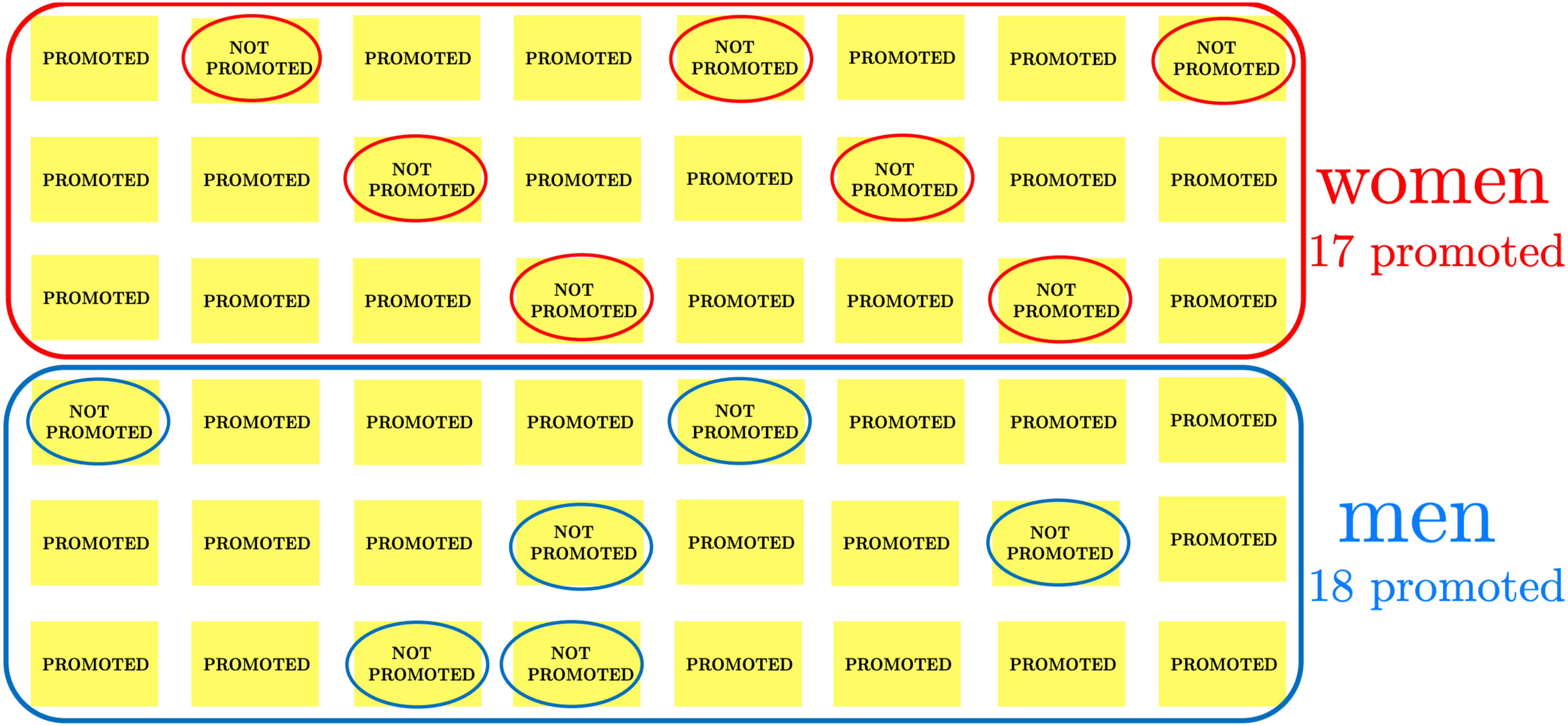
The percentage of women promoted $= \dfrac{14}{24} = 58\%$

So there is a 30% difference between men and women promoted. Could this have been due to chance?

We will define a **null hypothesis** which says that nothing is going on, and this difference is merely due to chance.

Then we will introduce an **alternative hypothesis** which says that something is going on, that discrimination is actually happening, and that the difference could not have occurred due to chance.

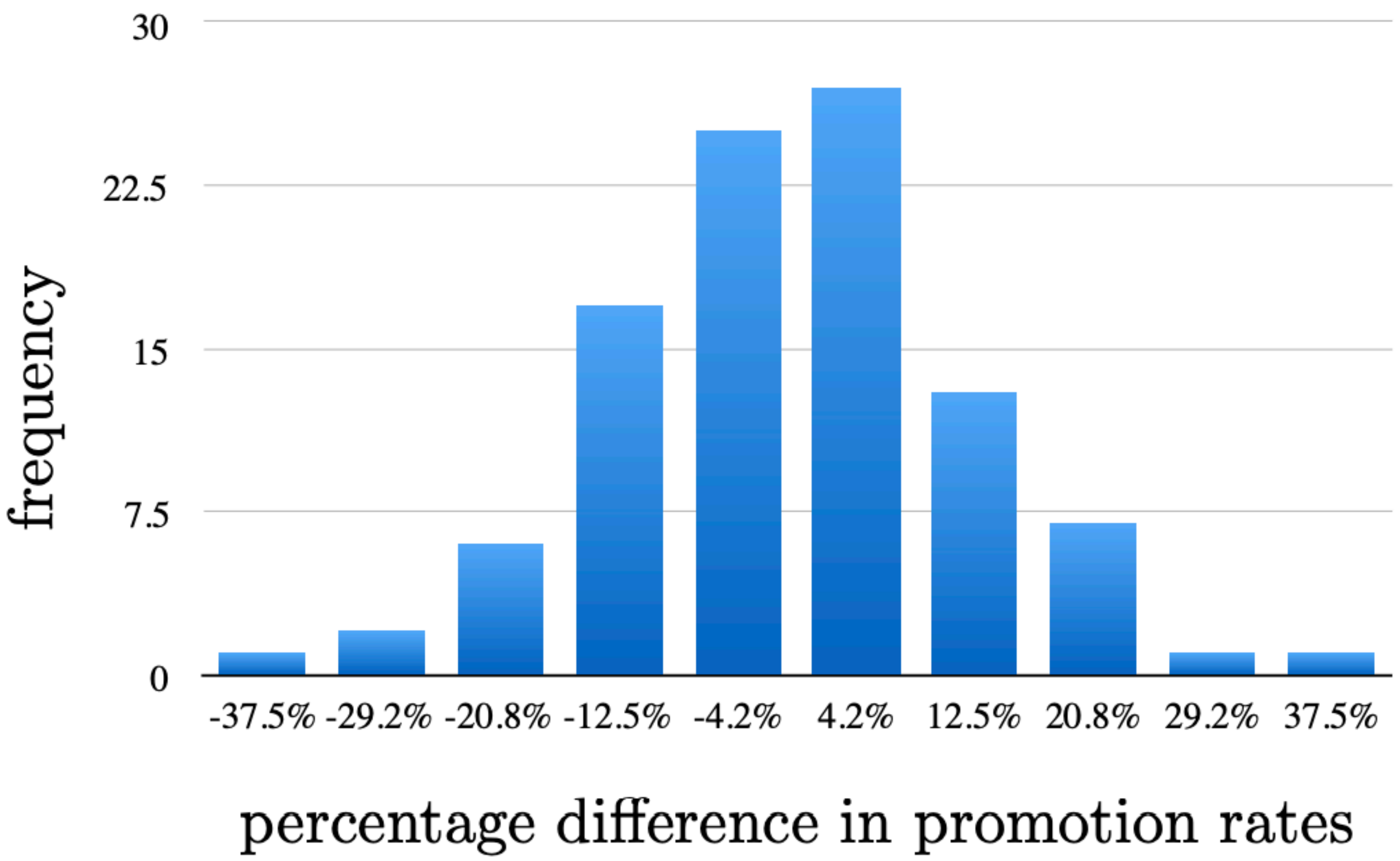
The null hypothesis is the status quo. We will stick to it unless we have evidence to favor the alternative hypothesis.



PROMOTED	NOT PROMOTED	PROMOTED	PROMOTED	NOT PROMOTED	PROMOTED	PROMOTED	NOT PROMOTED	
PROMOTED	PROMOTED	NOT PROMOTED	PROMOTED	PROMOTED	NOT PROMOTED	PROMOTED	PROMOTED	women 17 promoted
PROMOTED	PROMOTED	PROMOTED	NOT PROMOTED	PROMOTED	PROMOTED	NOT PROMOTED	PROMOTED	
NOT PROMOTED	PROMOTED	PROMOTED	PROMOTED	NOT PROMOTED	PROMOTED	PROMOTED	PROMOTED	
PROMOTED	PROMOTED	PROMOTED	NOT PROMOTED	PROMOTED	PROMOTED	NOT PROMOTED	PROMOTED	men 18 promoted
PROMOTED	PROMOTED	NOT PROMOTED	NOT PROMOTED	PROMOTED	PROMOTED	PROMOTED	PROMOTED	

Simulation: out of the 48 candidates (male and female), randomly draw 35 to get promoted.

We conduct this simulation 100 times, and each time we record the % difference between males and females promoted.



The probability to obtain a 30% or larger difference in promotion rates = 1%.

If the test results do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis.

If there is enough evidence, we reject the null hypothesis in favor of the alternative hypothesis

Since there is only 1% chance the results could have occurred by chance, we can reject the null hypothesis in favor of the alternative hypothesis, and we conclude that indeed there is gender discrimination.

Terminology:

H_0 : null hypothesis that there is no discrimination, i.e. each person is equally likely to get promoted.

H_A: alternative hypothesis, i.e., there is some sort of discrimination

p-value: probability of obtaining results that are at least as extreme as the observed results, if the null hypothesis were correct, i.e.,

$$p = P(\text{difference at least } 30\% | H_0) \approx 0.01$$

What value of the p - value is significant?

This is a value judgement and differs between fields, depending on how serious it would be to draw the wrong conclusion

$p = 0.05$ or $p = 0.01$ is common in biomedical research

$p = 0.0000003$ was used when discovering the Higgs boson

What value of the p - value is significant?

This is a value judgement and differs between fields, depending on how serious it would be to draw the wrong conclusion

$p = 0.05$ or $p = 0.01$ is common in biomedical research

$p = 0.0000003$ was used when discovering the Higgs boson

	H_0 is true	H_0 is false
Do not reject H_0	Correct inference	Type II error
Reject H_0	Type I error	Correct inference

Example 2:

Tom has two roommates: Ryan and Hugo. Every week, Tom draws a name out of a bucket to randomly select the roommates to take the trash out that week. Hugo suspects that Tom is cheating, so he starts keeping track of the draws, and he finds that out of 12 draws, Tom didn’t get picked even once!

H_0 : Tom is not cheating so each roommate gets picked 1/3 of the time

H_A : Tom is cheating

P(Tom not picked in a given draw $| H_0$) $=$

P(Tom not picked in 12 consecutive draws $| H_0$)

$=$

$p =$

P(Tom not picked in a given draw $|H_0) = \frac{\mathbf{2}}{\mathbf{3}}$

P(Tom not picked in 12 consecutive draws $|H_0)$

$=$

$p =$

P(Tom not picked in a given draw | H_0) $= \frac{2}{3}$

P(Tom not picked in 12 consecutive draws | H_0)

$$= \left(\frac{2}{3}\right)^{12} \approx 0.8\%$$

$p =$

P(Tom not picked in a given draw $| H_0) = \frac{2}{3}$

P(Tom not picked in 12 consecutive draws $| H_0)$

$$= \left(\frac{2}{3}\right)^{12} \approx 0.8\%$$

$$p = P(\text{Tom not picked in at least 12 draws}|H_0)$$

P(Tom not picked in a given draw $|H_0) = \frac{2}{3}$

P(Tom not picked in 12 consecutive draws $|H_0)$

$$= \left(\frac{2}{3}\right)^{12} \approx 0.8\%$$

$$\begin{aligned} p &= P(\text{Tom not picked in at least 12 draws}|H_0) \\ &\underset{\uparrow}{=} P(\text{Tom not picked in 12 draws}|H_0) \approx 0.008 \end{aligned}$$

because there were 12 draws in total

What value of the p - value is significant?

This is a value judgement and differs between fields, depending on how serious it would be to draw the wrong conclusion

$p = 0.05$ or $p = 0.01$ is common in biomedical research

$p = 0.0000003$ was used when discovering the Higgs boson

	H_0 is true	H_0 is false
Do not reject H_0	Correct inference	Type II error
Reject H_0	Type I error	Correct inference

Do not reject $\neq$ Accept!!!

Drawbacks of the Framework of Hypothesis Testing

- More inclined towards not rejecting the null hypothesis
- The result depends on the choice of null hypothesis