Corrigenda and Addenda

The preprint text and accepted text were compared using a script to produce the table of

corrigenda and addenda below.

Type	Text
	In the criminal justice system, defendants accused of a crime are presumed innocent until proven guilty. Statistical inference in any context is built on an analogous principle: The null hypothesis – often a hypothesis of "no difference" or "no effect" – is presumed true unless there is sufficient evidence against it. Of course, both criminal trials and statistical tests are susceptible to error. In some (hopefully very rare) cases, innocent people may be convicted. This first type of error – erroneously rejecting a true hypothesis – is mentioned frequently in introductory statistics classes, where students learn to choose a significance level (e.g. $\alpha=0.05$) to keep the type I error rate fairly low. However, there is another error that may occur: Guilty people may go free. This second type of error – erroneously failing to reject a false hypothesis – is often given only cursory treatment in introductory statistics courses, perhaps due to a lack of tidy formulas. In practice, type II error and its complement power, defined as 1 minus the probability of type II error, are important concerns. Researchers need to know the power of the statistical test to detect an effect – a particular difference between the null hypothesis and reality. If a study is under-powered, e.g. due to a small sample size, researchers may not be able to detect even a practically important difference because of the limitations of their study design. Because of its centrality and complexity, it is important that students have opportunities to explore statistical power and the factors that affect it in a manner that is concrete and accessible. The concepts of type I error, type II error, and power are all linked to the concept of sampling variability. Whether the null hypothesis is true or false, there is variability in the sample statistics, meaning it is not always clear whether an observed result is due to chance or a true effect. Thus, it is essential that students understand the variability that motivates questions of statistical power. The Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report recommends that understanding of variability be developed through active learning: "[Variability] is the essence of statistics as a discipline and not best understood by lecture. It must be experienced" (American Statistical Association 2005, 8). This call for active learning is echoed in the GAISE Pre-K-12 framework (Franklin et al. 2007). Chance, delMas, and Garfield (1999, 314)

	suggest using simulations "to give students a visual and more concrete understanding of sampling distributions," and they recommend students begin with physical simulations in a meaningful context before technological tools are introduced. The Common Core State Standards for Mathematics (NGACBP and CCSSO 2010) also include multiple standards related to simulation, introducing simulation of sample statistics in 7th grade (7.SP.2) and using simulation to determine statistical significance in high school (S-IC.5). In this article, we describe an activity with two parts – a physical simulation using spinners and a computer simulation using TinkerPlots software (Konold and Miller 2015) – through which students experience variability and explore the factors that affect statistical power. We developed this activity for an AP Statistics course that serves students in grades 10-12, but we believe it is equally appropriate for introductory statistics courses at the college level. In any high school or college course that introduces statistical significance tests, students can benefit from hands-on experiences with statistical power. Given certain sample sizes, effect sizes, and significance levels, the "guilty go free" more often than students may imagine!
Deletion	Introductory
Insertion	This first type of error – erroneously rejecting a true hypothesis – is mentioned frequently in introductory
Insertion	classes, where
Deletion	rejection rate for true hypotheses (i.e. the type
Insertion	type
Deletion	)
Deletion	effect
Deletion	that would be considered practically important in the study.
Insertion	– a particular difference between the null hypothesis and reality.
Deletion	
Insertion	(GAISE)
Insertion	This call for active learning is echoed in the GAISE Pre-K-12 framework (Franklin et al. 2007).
Deletion	." They also

Last Updated: 2021-10-02

Insertion	," and they
Insertion	also
Insertion	7th
Deletion	7
Insertion	(Konold and Miller 2015)
Insertion	high school or college
Insertion	Part 1: Physical Simulation
Insertion	We introduce the activity with a description of the alleged crime and the upcoming trial of Player A. In a 2-player trivia game, a spinner is used to decide which player gets the first shot at answering the question. Player A had access to the spinner before the game, and Player B suspects he may have tampered with it to get more chances at answering questions. As a group, test out the spinner to see if you can convict Player A of cheating. (Remember, he's innocent until proven guilty!) Make sure the judge can't see the spinner – just the outcomes of the spins. Students are then divided into groups of size four with each person choosing one the following roles: the All-Knowing Spinner, the Calculator Operator, the Graph Specialist, and the Honorable Judge. Descriptions of these roles are given in Table 1. Having each student in the group play a defined role serves two purposes: 1) every student can engage with the task and 2) the activity is expedited when it is clear how to divide the work. Group sizes could be modified if necessary by combining roles, provided that different students are the All-Knowing Spinner and the Honorable Judge for obvious reasons. Table 1. Roles for students in each simulation group.
Insertion	Role Description
Insertion	Role Description

	The All-Knowing Spinner This is the only member of the group that is allowed to see the spinner. This student spins the spinner and reports the outcome of each spin - Player A or B. The Calculator Operator This student continually updates the proportion of spins that result in Player A being selected. That is, the student calculates a new p A after each spin. The Graph Specialist This student receives the updated proportion from the Calculator Operator and updates a graph of the history of the estimate. This graph shows the spin number on the x axis and the current estimated proportion on the y axis. The Honorable Judge This student has the sole authority to decide whether there is enough evidence to convict Player A of cheating. It is imperative that the Honorable Judge not be allowed to see the spinner before they make their decision.
Insertion	The All-Knowing Spinner This is the only member of the group that is allowed to see the spinner. This student spins the spinner and reports the outcome of each spin - Player A or B.
Insertion	The Calculator Operator

	This student continually updates the proportion of spins that result in Player A being selected. That is, the student calculates a new p A after each spin.
Insertion	The Graph Specialist This student receives the updated proportion from the Calculator Operator and updates a graph of the history of the estimate. This graph shows the spin number on the x axis and the current estimated proportion on the y axis.
Insertion	The Honorable Judge This student has the sole authority to decide whether there is enough evidence to convict Player A of cheating. It is imperative that the Honorable Judge not be allowed to see the spinner before they make their decision.
Insertion	Multiple spinners are distributed to the groups, some more unfair than others. As shown in Figure 1, the spinners we use have, the proportion of the spinner allocated to Player A, as either 0.5 (a fair spinner), 0.6 (a moderately unfair spinner), or 0.75 (a severely unfair spinner). / Figure 1. Spinners representing p A =0.50, p A

=0.60, and p_A =0.75.

Groups are asked to complete at least 20 spins before the Honorable Judge makes a decision. Graph space is provided for up to 50 spins, and the judge is charged with deciding when enough evidence has been collected. Figure 2 displays the history of the estimate for a group that received a severely unfair spinner (p_A =0.75). Note the variability in

p_A

at the beginning of the simulation. It would have been imprudent to convict Player A after the first 5 or even 10 spins. However, as more spins were recorded, the estimate

p_A

began to stabilize near the true value p_A

. Some judges may have been convinced of Player A's guilt after 20 spins, but this particular judge demanded more evidence, asking her group to complete all 50 allowable spins.

Deletion	
	p *A*
Insertion	
Deletion	/
Insertion	/
Insertion	Figure 2. A plot displaying the history of *p* *A* . When the Honorable Judge is satisfied that sufficient evidence has been collected, he or she pronounces judgment on Player A: guilty or not guilty. At this point, the All-Knowing Spinner reveals the truth about Player A's guilt or innocence, and all groups come together for a class discussion of two key questions: What factors influenced whether or not Player A was convicted? What are the two possible errors you could make when judging Player A? The concept of power is introduced informally as students suggest factors that influence whether or not Player A is convicted. Certainly the number of spins (sample size) is important. Among our students, some judges were unable to make a decision after 20 spins but went on to convict Player A after seeing more evidence. Player A's conviction is also influenced by how unfair the spinner is (effect size). For example, groups given a severely unfair spinner often find Player A guilty after only a few spins, but groups with a moderately unfair spinner may fail to convict player A even after many spins. The conviction is also dependent on how much evidence the particular judge requires (analogous to significance level). For example, the judge whose decision was based on

<table>
<tr><td rowspan="1"></td><td>

the graph in Figure 2 above required airtight evidence to convict (a low significance level); in contrast, another judge convicted based on the estimate

p

A

= 0.55 after only 25 spins (a higher significance level).

The terminology of type I error, type II error, and power are introduced through discussion of the possible errors that could be made in the judgment of Player A. A type I error occurs if the judge convicts Player A when he is really innocent (analogous to rejecting the null hypothesis when it is true); a type II error occurs if the judge fails to convict Player A when he is really guilty (analogous to failing to reject the null hypothesis when it is false). For groups with moderately unfair spinners ($p_A = 0.6$), failure to convict Player A is somewhat likely. This illustrates that a Type II Error is not anyone's "fault" – everyone in a group committing a Type II Error did their jobs, but the difference was just too small to be detected with a limited number of spins. After introducing type I error and type II error, power can be defined formally as power = 1 – P(type II error). In this context, power is the probability of detecting that Player A is cheating (for a certain unfair spinner).

As students wonder how all of these moving pieces fit together, we move on to the next phase of the activity.

</td></tr>
<tr><td>Deletion</td><td>knowing</td></tr>
<tr><td>Insertion</td><td>Knowing</td></tr>
<tr><td>Deletion</td><td></td></tr>
<tr><td>Deletion</td><td></td></tr>
<tr><td>Deletion</td><td>

(

p

A

=0.6),

</td></tr>
<tr><td>Insertion</td><td>(pA = 0.6),</td></tr>
</table>

Deletion	
Deletion	power=1−P type II error .
Insertion	power = 1 – P(type II error).
Insertion	Part 2: Computer simulation
Insertion	Building on the class discussion of how sample size, effect size, and significance level affect power, we use computer simulations in TinkerPlots version 2.3.1 (Konold and Miller 2015) to investigate these conditions more systematically. Students work in pairs using a TinkerPlots file (tinyurl.com/RiggedGame) created to mimic the physical simulation from the first part of the activity. The model is shown in Figure 3. The TinkerPlots interface allows students to manipulate the spinner, the number of spins, and the significance level. Prior to this activity, our students had seen TinkerPlots used in demonstrations, but they had not used it themselves. After a thorough introduction, most students found the interface intuitive. /Figure 3. The TinkerPlots interface for this activity.
Insertion	version 2.3.1 (Konold and Miller 2015)
Deletion	Note that this model is based on a z test for a single proportion, so it may be helpful to review z tests before beginning the computer simulation.
Insertion	This TinkerPlots model is based on a one-sided z test for a single proportion. For each simulated trial of n spins, the TinkerPlots program computes a z statistic using the usual formula $z=$ p A

−

p

0

p

0

$(1-$

p

0

$)$

n

=

p

A

−0.5

0.5 (1−0.5)

	n
	where p0 is 0.5, because the null hypothesis assumes that Player A has not cheated and therefore the spinner probabilities are equal for both players. Students may confirm that the program works by calculating the first z statistic by hand using the formula.

When conducting a z test, we reject the null hypothesis whenever the z statistic is greater than the critical value, z*, that corresponds to our significance level, α. That is, the lower our significance level, the higher the z statistic must be to convince us that the spinner is unfair. We use one-sided critical values, because we assume that Player A would tamper with the spinner to get more (not fewer) chances at answering questions; that is, the alternative hypothesis is pA > 0.5. Before beginning the TinkerPlots simulation, students calculate these z* values using a table of standard normal cumulative probabilities, a calculator with statistical functions, or software. |
| Insertion | Simulating a large number of trials allows students to explore how the probability of conviction (power when Player A cheats and type I error when the spinner is fair) is related to sample size, effect size, and significance level. To guide their investigations, students complete a table like the one shown in Table 2 below by recording the estimated probability of conviction for each set of conditions. The probably of conviction is estimated by using TinkerPlots to find the proportion of simulated z-values that are greater than the critical value. (Note that the values shown in Table 2 are not exact probabilities but estimates based on simulations; students' results will differ slightly from these values due to random chance.) Students then use the information in the table to answer the following questions:

What happens to the probability of conviction as the significance level goes down?

Is this a bad thing, a good thing, or a trade-off? Explain.

What happens to the probability of conviction as the number of spins goes up? (Is the pattern different depending on which spinner you're testing?)

What happens to the probability of conviction as the spinner gets further from fair (as pA increases)?

If you want to be able to convict cheaters even when pA is only slightly bigger than 0.5, what will you have to do? |

Deletion	the
Insertion	a
Insertion	The probably of conviction is estimated by using TinkerPlots to find the proportion of simulated z-values that are greater than the critical value.
Deletion	p A
Insertion	pA
Deletion	p A
Insertion	pA
Insertion	
Insertion	Table 2. Hypothetical results for simulations of the conviction rate under various conditions Significance Level
Insertion	

Significance Level

α = 0.1

α = 0.05

α = 0.01

Spinner

Number of Spins

z

$*$

$=\ 1.28$

z

$*$

$=\ 1.64$

z

$*$

$=\ 2.33$

Insertion	α = 0.1 α = 0.05 α = 0.01
Deletion	α=0.1
Insertion	α = 0.1
Deletion	α=0.05
Insertion	α = 0.05
Deletion	α=0.01
Insertion	α = 0.01
Insertion	Spinner Number of Spins
Deletion	
Deletion	z $*$ $=\ 1.28$ z $*$

	$=\ 1.64$ z $*$ $=\ 2.33$
Deletion	
Deletion	$p=0.5$
Insertion	$p = 0.5$
Insertion	$/$ $n = 30$ 8% 5% 1% Type I error
Deletion	$n=30$
Insertion	$n = 30$
Insertion	 $n = 100$ 12% 4% 0%

Deletion	$n=100$
Insertion	n = 100
Deletion	$p=0.6$
Deletion	p = 0.6 / n = 30 44% 30% 3% Power n = 100 77% 62% 39% p = 0.75 / n = 30 95% 94%

	68% n = 100 100% 100% 99%
Insertion	p = 0.6
Insertion	/ n = 30 44% 30% 3% Power
Deletion	n=30
Insertion	n = 30
Insertion	 n = 100 77% 62%

Last Updated: 2021-10-02

	39%
Deletion	n=100
Insertion	n = 100
Deletion	p=0.75
Insertion	p = 0.75
Insertion	/ n = 30 95% 94% 68%
Deletion	n=30
Insertion	n = 30
Insertion	n = 100 100% 100% 99%
Deletion	n=100

Insertion	n = 100
Insertion	As they search for patterns in the table, students discover that the power increases as the significance level goes down, but the increased power comes at the cost of increased risk of type I error. In the context of "innocent until proven guilty," the question of significance level can lead to a rich discussion focused on the consequences of type I and type II errors. Students also discover that increases in sample size and effect size correspond to increased power; thus, large sample sizes are necessary if small effects are to be detected. Discussion and Next Steps This activity was designed for two 50-minute class periods or one 100-minute class period, and this time commitment is appreciable. Further, this activity does not introduce power using the superimposition of distributions for the null and alternative hypotheses, an abstract but traditional way of representing power. Given the importance of statistical power in practice and the difficulties we have experienced when introducing the topic more abstractly, we believe that using technology to develop the concept of power empirically is a worthy investment of time. The use of TinkerPlots presents its own challenges. The simulation as it is designed in TinkerPlots has some idiosyncrasies such as requiring the sample size to be updated in two locations (detailed instructions for using the TinkerPlots file are available as a supplementary file). Additionally, TinkerPlots is proprietary software with a fee associated with its use. Statistical power is an important but conceptually challenging topic for students in introductory statistics courses at both the high school and college level. To make this topic more accessible and engaging, this activity invites students to explore the factors that affect statistical power through an extension of the commonly used "innocent until proven guilty" metaphor for type I and type II error. Beginning with a hands-on simulation before moving on to its technological equivalent, students experience variability as it relates to statistical power and make connections between the results of physical simulations in the classroom and "long run" probabilities estimated from computer simulations In this activity, simulations are used to develop conceptual understanding of statistical power as well as introduce students to power studies similar to those undertaken by practicing statisticians. References

	American Statistical Association. 2005. Guidelines for Assessment and Instruction in Statistics Education: College Report. Alexandria, VA: American Statistical Association. Chance, Beth, Robert delMas, and Joan Garfield. 2004. "Reasoning about Sampling Distributions." In The Challenge of Developing Statistical Literacy, Reasoning and Thinking, edited by Dani Ben-Zvi and Joan Garfield, 295–323. Dordrecht, Netherlands: Kluwer Academic Publishers.
Deletion	
Deletion	; if this representation is to be introduced it would require additional time still. However, given
Insertion	. Given
Deletion	a full list of tips
Insertion	detailed instructions
Deletion	given elsewhere). Also
Insertion	available as a supplementary file). Additionally
Deletion	that has a licensing
Insertion	with a
Deletion	. To give users a choice, we are developing an R Shiny App that allows students to carry out the computer simulations using a free, web-based tool
Deletion	
Insertion	Franklin, Christine, Gary Kader, Denise Mewborn, Jerry Moreno, Roxy Peck, Mike Perry, and Richard Scheaffer. 2007. Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework. August 2005. American Statistical Association. Konold, Clifford, and Craig D. Miller. 2015. TinkerPlots (version 2.3.1). Learn Troop. http://www.tinkerplots.com/.
Deletion	Matematics
Insertion	Mathematics