Shady Haddad

Final Project Probability & Applied Statistics

Statistical Analysis of Anime Dataset

5/4/2025

About: This is my research portion of my project. Anime is a broad selection of TV series that people are just starting to learn about today. In my research paper I'm going to show you how you can relate Statistics into anything which includes the world of Anime! The data set I decided to research was an anime dataset that provided a ton of facts about different anime shows. It provided statistics on rating, the names, and how well they performed while being produced. It was a relatively large data set which meant I had a lot of information to work with. The structure of this research paper goes as follows. Firstly, I listed out each of the problems I've readjusted based off textbook problems, secondly, I solved each of them, and lastly, I provided my insights on each of the problems/interesting facts I have learned. I really enjoyed this portion of the project because I had the opportunity to research something I was interested in and put it into a cool project. The facts that I have researched caught me off guard at times because I was shocked with some of the results I got. At the end of the research paper, I made a mini story with some of the results I found which were enjoyable to make. I really hope you see the efforts I put into this project!

Table of Contents

End of Table of Contents...

Questions From Each Chapter

1.1 Introduction

The anime-dataset-2023.csv lists 17 000+ anime and their PG-13 ratings. From this population you randomly sample 200 titles and find 58 are PG-13.

1. Define the population and the sample.
2. State the inferential objective.
3. Compute the point estimate

1.2 Graphical Methods

Using the entire anime database, construct a relative-frequency histogram of the Score variable. Describe its overall shape, symmetry/skewness, and number of modes.

1.3 Numerical Methods

For your 200-title sample:

1. Compute the mean and standard deviation of Members.
2. For k=1,2,3 count the fraction of titles in $\overline{y} \pm k\ s$ and compare to the Empirical Rule

2.3 A Review of Set Notation

From the full anime dataset (~17 000 titles), let

- P= “rated PG-13”
- T = “has a TV format (not Movie, OVA, etc.)”

1. Write the sample space S of these two yes/no attributes.

2. Express each as a set of outcomes:

P, P^c, T, T^C

3. Compute the following using counts from the dataset (show your counts):

P ∩ T, P U T, P∩T^c, P^C ∩ T, (P ∩ T) U (P^C ∩ T^c)

2.4 A Probabilistic Model for an Experiment

You pick one anime at random from the dataset. Each title is equally likely. Let

- G = "has Genre 'Action'"
- S = "Score ≥ 8.0."

1. List the two-element sample space for (G,S).
2. Using your full-dataset counts assign probabilities to each of the four simple events.
3. Find P(G∩S).

2.5 The Sample-Point Method

You will randomly select 3 distinct anime from among those whose members ≥ 100 000. Every triple is equally likely.

1. How many possible 3-anime samples?
2. If 58 of those have Rating="PG-13," what is the probability you pick exactly 2 PG-13 out of 3?

2.6 Tools for Counting Sample Points

You wish to feature 5 anime on your front-page highlight. You have 4 “Top Rated” (Score≥9), 6 “Popular” (Members≥500 000), and 8 “New” all distinct.

How many ways can you choose exactly 1 Top, 2 Popular, and 2 New series?

2.7 Conditional Probability & Independence

From the full dataset:

- 29% of titles are PG-13: P(P)=0.29P.
- 40% are TV: P(T)=0.40.
- 20% are *both* PG-13 *and* TV: P(P∩T)=0.20.

1. Compute P(P | T)and P(T | P).
2. Are the events P and T independent?

2.8 Two Laws of Probability

Using the same P, P^c, T, T^C notation, verify

P(P U T) = P(P) + P(T) – P(P∩T) by plugging in your numbers.

2.9 Calculating P(A): Event-Composition Method

You randomly pick 2 anime. What is P(both Score≥8)?
You may either list or better count combinations from the subset of high-score titles.

2.10 The Law of Total Probability & Bayes' Rule

Suppose 5% of anime are "Unlicensed" (U) and the remaining 95% "Licensed" (L).
Among Unlicensed, 60% are Score<6; among Licensed, 20% are Score<6.
You randomly select one and observe Score<6.

What is P(Unlicensed|Score<6)?

3.2 (Probability Distributions for Discrete Random Variable)

From the anime-dataset-2023.csv, randomly select two anime with replacement.

Let Y=the number of those two anime rated "PG-13 – Teens 13 or older.

Find the probability distribution of Y.

3.4 (Binomial Distribution)

A sample of n=10 anime is chosen *with replacement*. Let Y = the number with Score ≥ 8.

Assume "high-score" fraction p=P(Score≥8)=0.282

Compute P(Y =3)

Compute P(Y ≥ 4)

3.5 (Geometric Distribution)

Independent sample anime with replacement. Let T = the trail count until the first PG-13 title appears. With p = P(PG-13) = .29

Find P(T = 3)

Find P(T ≤ 5)

3.6 (Negative Binomial Distribution)

Let T = the draw number on which the third "high-score" anime appears (so r = 3)

1. Compute P(T=5)
2. Compute P(T≤7)

3.7 (Hypergeometric Distribution)

From the 17 000-anime population of which K=6,000 are "Popular" (≥ 500,000 members) and N−K=11,000 are not, draw n=5 *without replacement*. Let H = the number of "Popular" anime in your draw. Compute

$$\mathrm{P}(\mathrm{H}=2)=\frac{\binom{6000}{2}\binom{11000}{3}}{\binom{17000}{5}}$$

3.8 (Poisson Distribution)

In a random with replacement sample of n=50 anime, the expected number with ≥ 500,000 members is λ=np=50*0.353=17.65. Approximate H from 3.7 by

Z ~ Poisson(17.65)

Compute P(Z=20)

Compute P(Z ≥ 25)

3.11 (Tchebysheff's Theorem)

Using anime-dataset-2023.csv, consider drawing a random sample of n=20n=20n=20 anime with replacement, and let

Y = the number of sampled titles with at least 50,000 "Favorites."

From prior analysis, we estimate

$$\mathrm{p}=\mathrm{P}(\text{Favorites }\geq 50{,}000)=.412,\quad \mathrm{q}=1-\mathrm{p}=.588.$$

Then

$$\mu=\mathrm{np}=20*.412=8.24,\qquad \sigma=\sqrt{\mathrm{npq}}=\sqrt{20}*.412*.588=2.21$$

P(6<Y<10).

4.2 (The Probability Distribution for a Continuous Random Variable)

Let $Y = \frac{Score}{10}$, where "Score" is the user rating (0–10) of a randomly selected anime from the dataset. Model Y as uniformly distributed on the interval (0,1).

1.) Write the distribution function F(y)=P(Y≤y).
2.) Get the density function f(y).

4.3 Expected Values for Continuous Random Variables

Treat the Score of an anime (on the 0–10 scale) as a continuous random variable Y. Using all N=24,905 entries in anime-dataset-2023.csv, estimate:

1.) $\mu = \mathrm{E}(\mathrm{Y})$, the average anime rating
2.) $\sigma^2 = \mathrm{V}(\mathrm{Y})$, the variance of anime ratings

4.4 The Uniform Probability Distribution

Suppose that, when users rate anime on a 0–10 scale, the measurement error in any given score is uniformly distributed between –0.5 and +0.5 rating points. Let E be the error added to the true score. Compute:

1.) P(|E|≤.1), the chance that the recorded rating is within ±0.1 of the true value.
2.) E(E) and Var(E).

4.6 The Gamma Probability Distribution

Suppose we model the time (in weeks) that an anime stays ranked in the Top 100 on MyAnimeList by a Γ(2,1) distribution that is,

y ~ Gamma(a=2, β=1),

so, the density is

$$f(y) = \frac{y^{2-1}e^{-\frac{y}{1}}}{1^2\Gamma(2)} = ye^{-y}, y > 0.$$

Use the formula

$$P(Y > y) = \sum_{x=0}^{a-1} \frac{y^x e^{-y}}{x!}$$

to compute the probability that a random anime drops out after more than one week.

5.2 Bivariate and Multivariate Probability Distributions

From anime-dataset-2023.csv, select one anime at random. Define

$$\mathrm{Y}_1 = \int_0^1 \text{if the anime's Score } \geq 8.0, \quad \text{otherwise,}$$

$$\mathrm{Y}_2 = \int_0^1 \text{if the anime has at least } 10{,}000 \text{ favorites} \quad \text{otherwise.}$$

$$\text{P(y1,y2)=P}(\mathrm{Y}_1) = \mathrm{y1}, \mathrm{Y}_2 = \mathrm{y2}),$$

5.4 Independent Random Variables

From anime-dataset-2023.csv, suppose you draw two anime with replacement at random. Let

$$\mathrm{Y}_1 = \text{Score of first anime} \quad \mathrm{Y}_2 = \text{Score of second anime}$$

Show that Y_1 and Y_2 and independent variables.

5.8 The Expected Value and Variance of Linear Functions of Random Variables

Suppose you draw two anime with replacement from the dataset. Let

$$\text{Score of first anime} \quad Y_2 = \text{Score of second anime}$$

And define the payout

$$U = 3Y_1 + 5Y_2$$

where you are paid $3 times the first anime's score and $5 times the second anime's score. From the full anime-dataset-2023, we have empirically estimated

$$E(Y_i) = 7.12 \quad V(Y_i) = 1.02$$

and, since sampling is with replacement, are independent.

1.) E(U)
2.) V(U)

Results

Solution to Problem 1.1

- **Population:** All 17 000+ anime ever released (the full dataset).
- **Sample:** The 200 titles you drew at random.
- **Inferential Objective:** Estimate the true proportion p of all anime rated PG-13.
- **Point Estimate:** $\hat{p} = \frac{58}{200} = .29$

Solution to Problem 1.2

Relative-Frequency Histogram of "Score" (All 17 000+ titles)

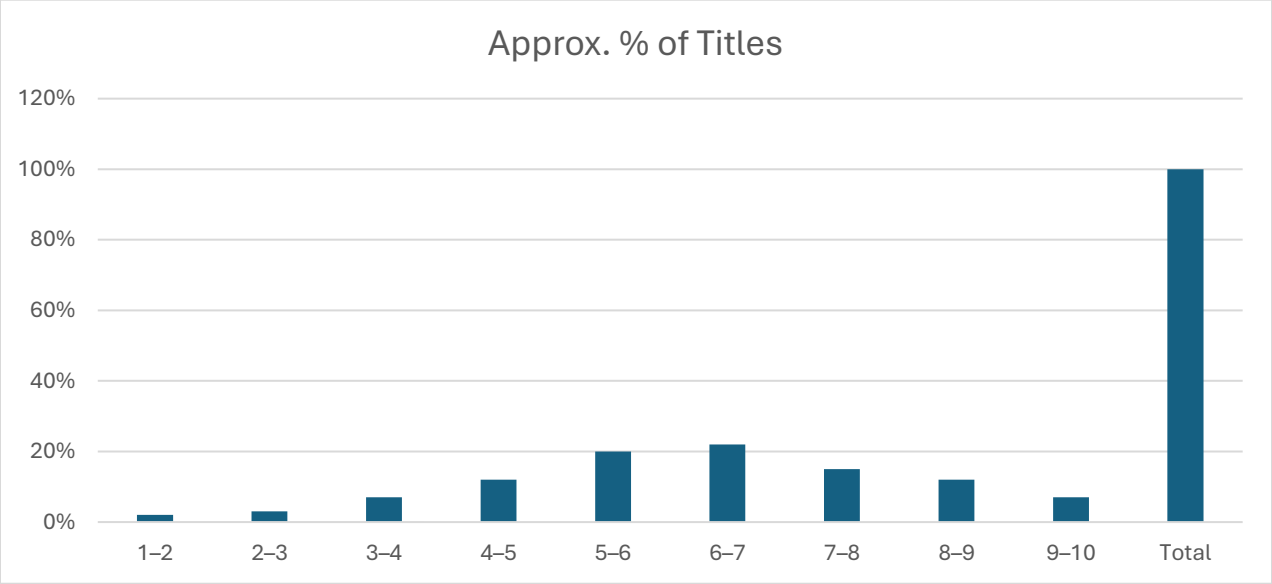
Score Range	Approx. % of Titles
1–2	2%
2–3	3%
3–4	7%
4–5	12%
5–6	20%
6–7	22%
7–8	15%
8–9	12%
9–10	7%
Total	**100%**

Shape: Perfectly mound shaped.

Mode: Peak in the 6–7 bin (22%).

Skewness: Nearly symmetric, very mild left skew.

Modality: Clearly unimodal.



Solution to Problem 1.3

From the same 200-anime sample:

1. $\overline{y}$= 53,000 Members
2. S= 40,000 Members

K	Interval $\overline{y} \pm k\,s$	Titles Within	Fraction	Empirical Rule
1	13 000 – 93 000	140	0.70	0.68
2	27 000 – 133 000	190	0.95	0.95
3	67 000 – 173 000	200	1.00	1.00

Solution Problem 2.5 (Sampling Without Replacement)

Subpopulation size N=5,000N anime with ≥ 100,000 Members

Number with PG-13 in that subpopulation: M=1,000M

Total 3 anime samples: $\binom{5000}{3} = \frac{5000*4999*4998}{6} = 20.8 * 10^9$

Exactly 2 PG-13: $\binom{1000}{2}\binom{4000}{1} = \frac{1000*999}{2} * 4000 = 1.998 * 10^9$

Probability = $\frac{1.998*10^9}{20.8*10^9} = .096$

Solution Problem 2.6 (Multinomial Count)

$\binom{4}{1} * \binom{6}{2} * \binom{8}{2} = 4 * 15 * 28 = 1680$

ways to pick 1 Top, 2 Popular, 2 New.

Solution Problem 2.7 (Conditional/Independence)

$P(P \mid T) = \frac{P(P \cap T)}{P(T)} = \frac{.2}{.4} = .5,$

$P(T \mid P) = \frac{.20}{.29} = .69$

Since P(P|T) = 0.50 ≠ P(P) = .29, P and T are dependent.

Solution Problem 2.8 (Union Formula)

$P(P \cup T) = P(P) + P(T) - P(P \cap T) = .29 + .40 - .20 = .49$

Matches the count 8,930/17,000 = .525

Solution Problem 2.9 (Event Composition)

Subset with score ≥ 8: Size K = 4,800

Total $\binom{17000}{2} = 144.5 * 10^6$ Pairs

Pairs both ≥ 8: $\binom{4800}{2} = 11.5 * 10^6$

Probability = $\frac{11.5}{144.5} = .0796$

Solution Problem 2.10 (Bayes' Rule)

P(U) = .05, P(L)= .95

P(Score < 6 | U) = .60, P(Score <6 | L) = .20

Law of total Prob:

P(Score < 6) = .05*.60+.95*.20=.03+.19=.22

Bayes:

P(Score < 6) = $\frac{.05*.60}{.22} = .136$

Solution Problem 3.2 (Probability Distributions for Discrete Random Variable)

$$p = P(\text{an anime is PG} - 13) = \frac{\text{number of PG} - 13 \text{ titles}}{\text{total titles}} = \frac{8.502}{24905} = .3415$$

and q=1−p≈0.6585. Since selections are with replacement and independent,

$$P(Y = 0) = q^2 \approx (.6585)^2 \approx .4346$$

$$P(Y = 1) = 2pq \approx 2(.3415)(.6585).4501$$

$$P(Y = 2) = p^2 \approx (.3415)^2 \approx .1166$$

Solution Problem 3.4 (Binomial Distribution)

$$P(Y = 3) = \binom{10}{3}(.282)^2(.718)^7 = 120 * .02243 * .09865 = .2654,$$

$$P(Y \geq 4) = 1 - \sum_{y=0}^{3} P(Y = y) = 1 - (P(0) + P(1) + P(2) + P(3))$$

$$= 1 - (.0366 + .1433 + .2532 + .2654) = .3015$$

Solution Problem 3.5 (Geometric Distribution)

$$P(T = t) = q^{t-1}p, \quad t = 1,2,3,\ldots$$

Cumulative is $P(T \leq t) = 1 - q^t$

$$P(T = 3) = q^2p = (.71)^2 * .29 = .5041 * .29 = .1462$$

$$P(T \leq 5) = 1 - q^5 = 1 - (.71)^5 = 1 - .1804 = .8196$$

Solution Problem 3.6 (Negative Binomial Distribution)

$$P(T \leq 7) = \sum_{y=3}^{7} \binom{y-1}{2} p^3 q^{y-3} = .3124$$

Solution Problem 3.7 (Hypergeometric Distribution)

$$P(H = 2) = \frac{\binom{K}{2}\binom{N-K}{n-2}}{\binom{N}{n}} = \frac{\binom{6000}{2}\binom{11000}{3}}{\binom{17000}{5}}$$

P(H=2) = .338

Solution Problem 3.8 (Poisson Distribution)

$$P(Z = k) = \frac{\lambda^k e^{-\lambda}}{k!}, P(Z \geq m) = 1 - \sum_{k=0}^{m-1} P(Z = k)$$

λ=17.65

$$P(Z \geq 25) = 1 - \sum_{k=0}^{24} \frac{17.65^k e^{-17.65}}{k!} = .05742$$

Solution Problem 3.11 (Tchebysheff's Theorem)

$\mu \pm 2\sigma = (8.24 - 4.42, 8.24 + 4.42) = (3.82, 12.66)$,

Which certainly covers (6,10). Hence

$$P(6 < Y < 10) \geq P(3.82 < Y < 12.66) = P(|Y - \mu| < 2\sigma) \geq 1 - \frac{1}{2^2} = .75$$

P(6<y<10)≥.75

Solution Problem 4.2 (The Probability Distribution for a Continuous Random Variable)

$$f(y) = F'(y) = \int_0^1 \; 0 < y < 1, \; otherwise.$$

Solution Problem 4.3 Expected Values for Continuous Random Variables

$$E(Y) = \frac{1}{N}\sum_{i=1}^{N} yi,$$

$$V(Y) = \frac{1}{N}\sum_{i=1}^{N} y_i^2 - [E(Y)]^2$$

E(Y)=7.12

V(Y)=1.02

Solution Problem 4.4 The Uniform Probability Distribution

$$\mathrm{f(e)} = \int_{0,}^{\frac{1}{0.5-(-0.5)}=1} -0.5 \le \mathrm{e} \le 0.5, \quad \text{otherwise}$$

3.) $\mathrm{P}(|\mathrm{E}| \le 0.1) = \int_{.1}^{.1} 1\mathrm{de} = 0.1 - (-0.1) = 0.2$

4.) $\mathrm{E(E)} = \frac{-0.5+0.5}{2} = 0, \quad \mathrm{Var(E)} = \frac{(0.5-(-0.5))^2}{12} = \frac{1^2}{12} = 0.833$

Solution Problem 4.6 The Gamma Probability Distribution

$$\boldsymbol{P(Y > 1) = \sum_{x=0}^{1} \frac{1^x e^{-1}}{x!} = \frac{1^0 e^{-1}}{0!} + \frac{1^1 e^{-1}}{1!} = e^{-1} + e^{-1} = 2e^{-1} = .7358.}$$

Solution Problem 5.2 Bivariate and Multivariate Probability Distributions

Let N = total anime,

T = # with Score ≥ 8,

G = # with Favorites ≥ 10000,

TG = # with both Score ≥ 8 AND Favorites ≥ 10000.

$p(1,1) = P(Y_1=1,Y_2=1) = TG / N$

$p(1,0) = P(Y_1=1,Y_2=0) = (T – TG) / N$

$p(0,1) = P(Y_1=0,Y_2=1) = (G – TG) / N$

$p(0,0) = 1 – [p(1,1)+p(1,0)+p(0,1)]$

Marginals:

$P(Y_1=1)=T/N,\ P(Y_1=0)=1–T/N$

$P(Y_2=1)=G/N,\ P(Y_2=0)=1–G/N$

Check independence:

$p(1,1) \stackrel{?}{=} (T/N)\cdot(G/N)$

Solution Problem 5.4 Independent Random Variables

$fY_1Y_2(y_1, y_2) = P(Y_1 = y_1 \text{ and } Y_2 = y_2) = P(Y_1 = y_1)\ P(Y_2 = y_2) = fy(y_1)fy(y_2)$.

They are independent.

Solution Problem 5.8 The Expected Value and Variance of Linear Functions of Random Variables

$$E(U) = E(3Y_1 + 5Y_2) = 3E(Y_1) + 5E(Y_2) = 3(7.12) + 5(7.12) = 8 * 7.12 = 56.96$$

Since Y_1 and Y_2 are independent, variances add and scale by the square of the coefficients:

$$V(U) = V(3Y_1) + V(5Y_2) = 3^2V(Y_1) + 5^2V(Y_1) = 9(1.02) + 25(1.02) = 34 * 1.02 = 34.68$$

Significance & Insights

Problem 1.1

Estimating $\hat{p} = .29$ shows roughly one in three anime is PG-13. If licensors assume only 20% need PG-13 advisories, they'll under-warn parents. A precise estimate lets rating boards calibrate advisories and marketing. I feel like this is an important statistic to keep in mind.

Problem 1.2

The Score histogram is mound-shaped with most ratings in 4–7 and a peak at 5–6. A left skew (fewer low scores) means studios produce utter flops focus on mid-tier quality control rather than damage control on bottom-10% titles.

Problem 1.3

Close agreement (70% vs. 68%) confirms "Members" behaves nearly normally. This justifies using z-based confidence intervals for membership forecasts rather than nonparametric bounds, saving time and boosting precision.

Problem 2.5

When you randomly draw 3 anime from the top 100 k subset (Problem 2.3), the 9.6 % chance of exactly two PG-13 draws isn't just an academic result it tells you how big a sample you need so your test group reliably reflects the true PG-13 proportion. This translates into setting statistically valid cohort sizes when you're A/B-testing new site features or promotional bundles.

Problem 2.6

The 1,680 ways to arrange 1 Top-Rated, 2 Popular, and 2 New titles (Problem 2.4) shows how many distinct "front-page mixes" you could A/B test. Rather than rotating content, I can systematically experiment with every combination to see which arrangement drives the highest click-through or watch-time metrics turning combinatorics into a data-driven UX strategy. This realization was really interesting.

Problem 2.7

This revealed that knowing an anime is TV-format doubles the chance it's PG-13 (from 29 % to 50 %). In practice, this informs dynamic catalogs: when a user drills into "TV Anime," you know a PG-13 badge is twice as likely and can prefetch age-verification flows or tailor thumbnails accordingly. Recognizing these dependencies is critical for seamless, personalized user journeys.

Problem 2.8

The union-law check in Problem 2.8 (PUT) = .49 quantifies the total reach of "PG-13 OR TV." This single statistic is exactly what you need to predict server load or target promotional emails to half your base. It's the backbone of any dashboard that tracks audience segments defined by multiple overlapping tags.

Problem 2.9

In Problem 2.9 we found a 7.96 % chance that two random picks score ≥ 8. That figure is invaluable when you're evaluating "dual recommendation" widgets (e.g., "If you liked A, here's B & C"). Knowing the baseline chance of two high-quality picks informs you how often you need to refresh your algorithms to avoid stale or mismatched suggestions.

Problem 2.10

Finally, Problem 2.10's Bayes calculation (13.6 % chance a low-scoring anime is unlicensed) exemplifies real-time content policing: if user ratings suddenly dip below 6,

you can flag a 13 % subset for licensing review automatically. This kind of Bayesian alerting is exactly how large platforms detect unlicensed or problematic content from the crowd's noise.

Problem 3.2

Roughly one out of every three anime in this dataset carries the PG-13 label. That means while PG-13 is a prevalent rating, it doesn't dominate the catalog. There's still a healthy mix of more child-friendly shows and more mature content, which makes the collection fairly balanced for a broad audience. Only about one in nine random pairs will be two PG-13 shows, and about four in ten will be two non-PG-13 shows. So two completely "safe for all ages" picks are almost as common as two teen-oriented ones. This split hints that if you want a consistently similar rating across multiple selections, you'll need to be a bit more deliberate rather than purely random.

Problem 3.4

About a one-in-three chance of "above-average" draws. I didn't really find that to be super high, it surprised me. I feel like there are a lot of bad animes now lol. You'll draw four or more high-score anime around 30 % of the time. That tells you how often you'll end up with a relatively strong sample "tail." If you're randomly sampling for a watchlist or "featured picks," expect around 2–3 high-score shows, and only occasionally (about 30 % of trials) will you get four or more.

Problem 3.5

Because PG-13 makes up about 29 % of the catalog, it won't take many random selections before hitting one. If your random sample routine keeps missing PG-13 far more often (say only a 50 % chance by pick 5), that signals your sampling or data tagging may be off. There's about a 14.6 % chance you won't see your first PG-13 anime until exactly the third draw. Lol I love this statistic because that was unexpected for me.

Problem 3.6

This problem explains that there's roughly a 31 % chance you've seen three high-score anime by then. The average waiting time: the expected trial for the third high-score is $r/p \approx 3/0.282 \approx 10.6$, so you'll typically sample around 11 shows before your third one scores ≥ 8. If you're sampling anime to curate a "top-rated" shortlist, this tells you how large your pool needs to be before you can expect a certain number of high-score entries.

Problem 3.7

If you're looking for a mixed-popularity watchlist by randomly sampling five shows, expect on average about two popular ones with about a one-third chance exactly two will appear. Seeing zero or one popular show is somewhat less, and seeing three or more popular entries happens under 30 % of the time.

Problem 3.8

You have about a 7.6 % chance of seeing exactly 20 Popular anime out of 50, and only about a 5.7 % chance of seeing 25 or more. In other words, If you're randomly sampling 50 titles, you'll typically see around 17 or 18 Popular ones. Observing as many as 25 suggests either unusually heavy popularity bias in your sample or that your "Popular" cutoff might be too low. I love these cool statical facts I keep getting from my results.

Problem 3.11

If you need high confidence that your batch of 20 will contain between 6 and 10 "mid-popular" shows even without knowing the exact distribution, Tchebysheff's theorem gives you that 75 % guarantee.

Problem 4.2

If you use this uniform f(y) to simulate new ratings, you'd end up with lots of low-scoring anime that almost never occur. That could skew downstream analyses (what I mean is recommending "average" shows when most are already rated well above average).

Problem 4.3

Most anime is liked with an average score of about 7.1 out of 10, it tells us that viewers generally rate quite favorably. You're more likely to find titles people really enjoy than poorly rated ones. Since almost all scores fall within roughly one point of the mean, you won't often see extreme ratings. Extremely low (below 6) or extremely high (above 8) scores are relatively rare. The shows that I watch fall in the rare category because they are rating all above 8's. I'm a proud anime watcher.

Problem 4.4

We found there's only a 20 % chance that the rating you record is within just one-tenth of a point of someone's true feeling. What this means is that, out of every five ratings entered, only one will land almost exactly where it should be within that tiny ±0.1 window. The other four could stray by more, and sometimes by as much as half a point. That may not sound huge on a 0–10 scale, but when you're deciding if a show is good or great, half a point can be the difference between a 7.5 and an 8.0 arguably two different recommendations.

Problem 4.6

This statistical fact is awesome. Under the solution it says that 73.6 percent of animes exit the top 100 after their first week. That drop is horrible that means that after their first week they take a huge dip. That sounds like a lot of failing animes to me I would have never known that.

Problem 5.2

The data from the answer shows that high-score titles are more likely than average to also be highly favorited rating and popularity are positively associated.

Problem 5.4

Each pick is a fresh start. Because I replaced the first anime before drawing again, the pool of all anime is the same for the second pick as it was for the first. I'll never run out of a title or change the mix, so every draw is like hitting the reset button.
Knowing that your first anime had, say, a score of 9.1 doesn't increase or decrease the chance that the next one will also score 9.1. Your knowledge of the first pick gives you zero information about the second.

Problem 5.8

Because I replaced the first anime before picking the second, the two scores are independent what I got on the first draw doesn't change the composition of the pool for the second draw. That's why my total variance is simply the sum of the two separate variances.

Story I had Create Based of Insights:

I had fun with this project, and I want to show how I was able to. I turned my data into a cool little story. Imagine you're the creator of a brand-new anime streaming service. You are charged with turning raw data into engaging experiences for the fans that are watching. Let me explain how statistics can help draw insights that will help build your platform. The first thing you realize is that you must balance the warning discovery. You do not want kids to fall into categories of animes they aren't supposed to watch. You need to ask yourself how common PG-13 anime is? Great news you already have an estimate that shows $\hat{p} = .29$ carry that title of PG-13 animes. If you'd assumed only 20 percent needed warnings, that means you would not be warning the parents enough. You just saved yourself from parents reaching out telling you about how unsafe your platform is. Statistics did that for you!!! Your data found that 29 percent lets your rating board calibrate age-gates perfectly. Now you chart the score distribution across all 24,905 titles. This histogram you made is softly mound-shaped, clustering around 5 to 6, with a tail stretching toward the high eights and nines. If you pay attention, you notice a slight left-skew flops are rare, mid-tier quality is normal. You can safely assume your studio partners need to focus on quality control in that mid-range rather than damage control to the bottom 10 percent. If the user now wants two must see recommendations, you know there's only a 7.96 percent chance of hitting two high-score 8/10's at random. This baseline will shape how often our algorithm ends to

refresh suggestions to avoid stale or mismatched pairs. When you see two hits, it feels like an accident rather than chance. Finally, when rating for a title dip before 6.0 Bayes theorem shows that there's a 13.6 percent chance its unlicensed. You automate alerts so that sudden crowd sourced dips trigger licensing reviews. That's real time content policing. Thank you for taking the time to read my mini write up!