**Q2 [11% of total mark]**

1. Use baseR only (i.e., only things you were taught before Week 3) to keep only the people who won 1st or 2nd *and* achieved an audience rating of 8 or more. You don’t need to assign the result to any tibble (and don’t write over the existing d!) but your output should look like the screenshot below when it is knitted. (Don’t worry if the order of the rows/columns is different, but there should be the same number of rows and columns and they should have the same values).

d(judge c(1,2) & audience >= 8)

d["judge" c(1,2) & "audience" >= 8,]

top\_judge <- d$judge [c(1,2)]

top\_audience <- d$audience >= 8

d[c(top\_judge, top\_audience)]

is.na

1. Use function(s) from tidyverse that you were taught in Week 3 to accomplish the same task as in part (a): keep only the people who won 1st or 2nd *and* achieved an audience rating of 8 or more. As before, you don’t need to assign the result to any tibble (and don’t write over the existing d!). Your output should look like the screenshot below when it is knitted. (Don’t worry if the order of the rows/columns is different, but there should be the same number of rows and columns and they should have the same values).

d %>%

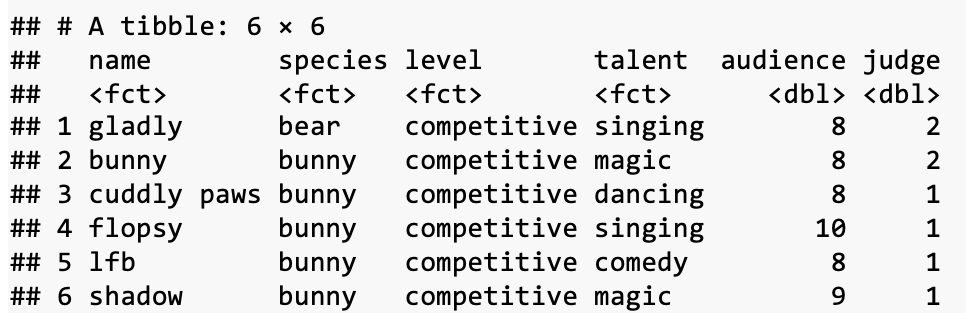
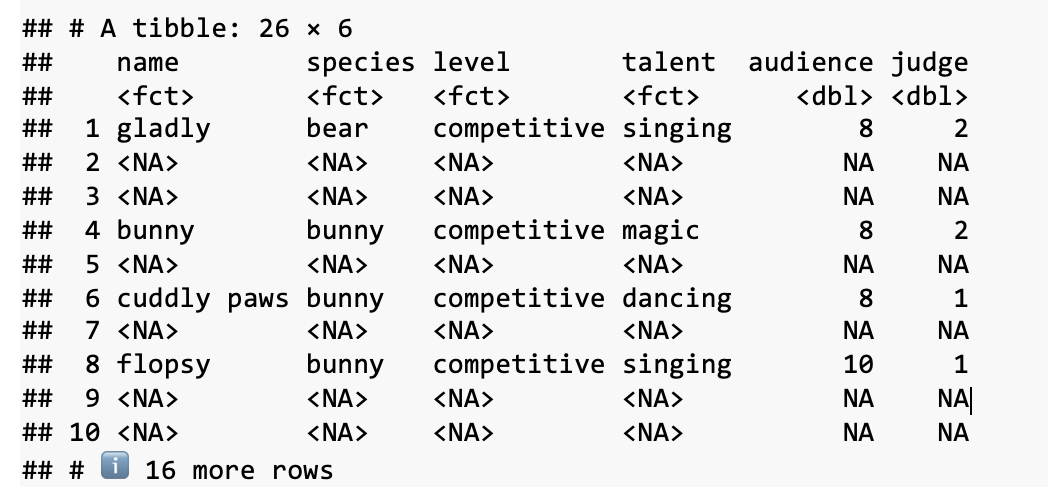
filter((judge == 1 | judge == 2) & audience >= 8)

1. You will notice that (b) and (a) do not match. Why? Answer in terms of what exactly the relevant part of baseR code is doing and how that is different from what exactly the relevant tidyverse code is doing. Note that you don’t need to discuss all of the components of your code, just the parts that are relevant to explaining the difference between (a) and (b).

The filter-function removes rows, and only extract the ones that match our condition. While the use of the logic functions shows the table in it's whole, but only displays the values that match the criterion put forward through the logic function. In this case the data has to be 1 or 2 in the judge-column, and 8 or above in the audience-column.

*[Suggested word count: 100]*

(d) Use baseR only (i.e., only things you were taught before Week 3) to create output that matches the screenshot in (b). As before you don’t need to assign the result to any tibble, just make sure that the output when knitted looks like (b). (Don’t worry if the order of the rows/columns is different).



Page 4 of 11

**Q3 [12% of total mark]**

(a) Use a single tidyverse function you were taught to remove the *judge* and *audience* columns from d and assign the result to a new tibble called dshort. Make sure that the top rows of dshort are visible in the knitted Markdown.

(b) Use tidyverse function(s) you were taught in Week 3 to transform dshort so that it looks like the tibble in the screenshot below. (Don’t worry if the order of the rows/columns is different, but there should be the same number of rows and columns with the same values). Assign the result to a new tibble called d2. Make sure that the top rows of d2 are visible in your knitted Markdown.

d2 <- dshort%>%

select(name, species)

mutate(competitive = level == "competitive" & talent, NA\_character\_)

mutate(fun = if\_else(level == "fun" & talent, NA\_character\_)

head(d2)

d2 <- dshort %>%

mutate(

competitive = if\_else(level == "competitive", talent, NA\_character\_),

fun = if\_else(level == "fun", talent, NA\_character\_)

) %>%

select(-c(level, talent)) %>%

d2 <- dshort %>%

group\_by(name, species, competitive, fun) %>%

mutate(

competitive = if\_else(level == "competitive", talent, NA\_character\_),

fun = if\_else(level == "fun", talent, NA\_character\_)

) %>%

ungroup() %>%

select(-c(level, talent)) %>%

print(d2)

11.35, 13 april:

d2 <- dshort %>%

group\_by(name, species, competitive, fun) %>%

mutate(

competitive = if\_else(level == "competitive", talent, NA\_character\_),

fun = if\_else(level == "fun", talent, NA\_character\_)

) %>%

ungroup() %>%

select(-c(level, talent)) %>%

print(d2)

Riktig kode:

d2 <- dshort %>%

group\_by(name, species) %>%

mutate(

competitive = if\_else(level == "competitive", talent, NA\_character\_),

fun = if\_else(level == "fun", talent, NA\_character\_)

) %>%

summarise(

competitive = first(na.omit(competitive)),

fun = first(na.omit(fun)), .groups = 'drop'

) %>%

ungroup() %>%

print(d2)

(c) Why did we have you perform the transformation in (b) using dshort instead of d? In other words, what happens if you were to do it on d, and why does this happen? You do not need to show any code or output to get full marks on this question but you can if you want to. If you do, be sure to refer to the code or output in your answer so it is clear why/how it is relevant.

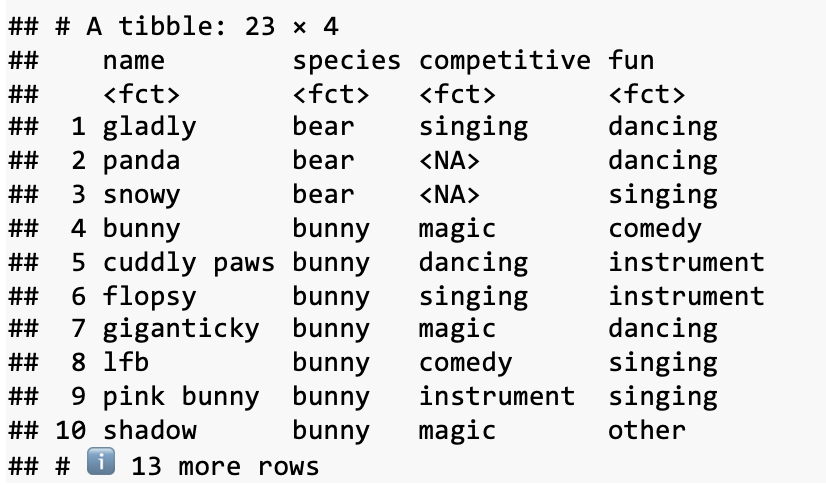
*[Suggested word count: 100]*

(d) Use your d2 tibble to determine if anybody broke either of the two rules of the talent show that are explained in the description for *level* in Table 1. For each rule, you should include code that identifies individuals that broke this rule – don’t just look at the tibble manually to find them. In your answer, be sure to list everyone who broke a rule along with what rule(s) they broke. If you did not succeed in creating d2 in part (b), you can use the tibble called d3b that has already been loaded for you.

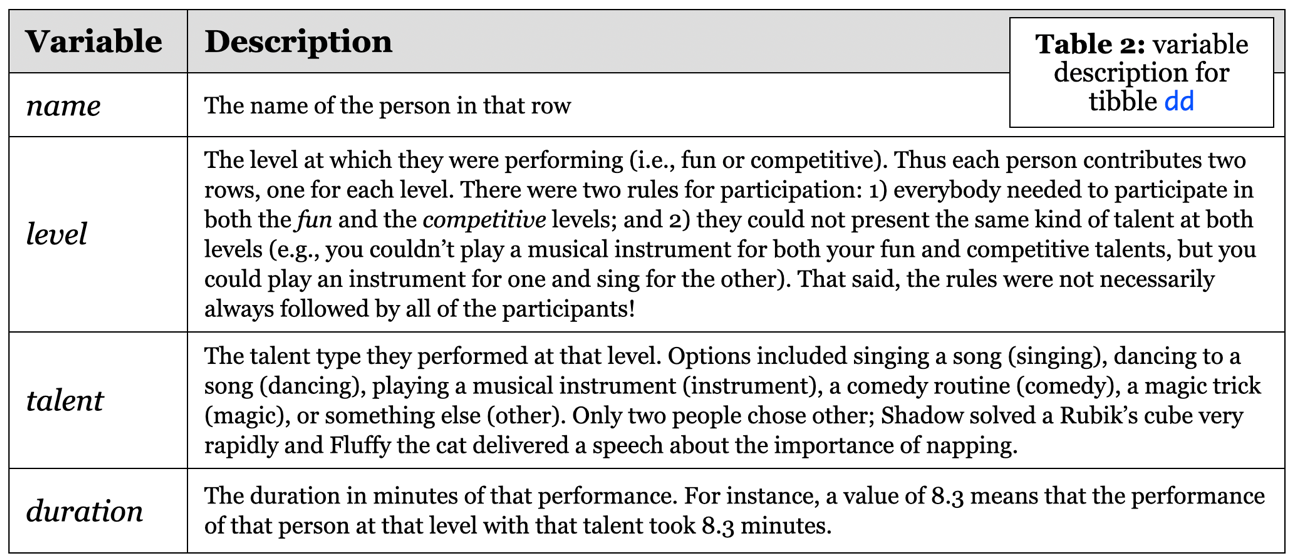
**Q4 [7% of total mark]**

(a) Change d so that the order of the *name* variable in it is alphabetical. Make sure that the top rows of d are visible in the knitted Markdown.

(b) One of the tibbles that has already been loaded for you is called dd. It contains the same data as d in the columns *name*, *level*, and *talent* (i.e., the same people and performances) but contains a new variable. A full explanation of the variables in dd is shown in Table 2.



Page 5 of 11

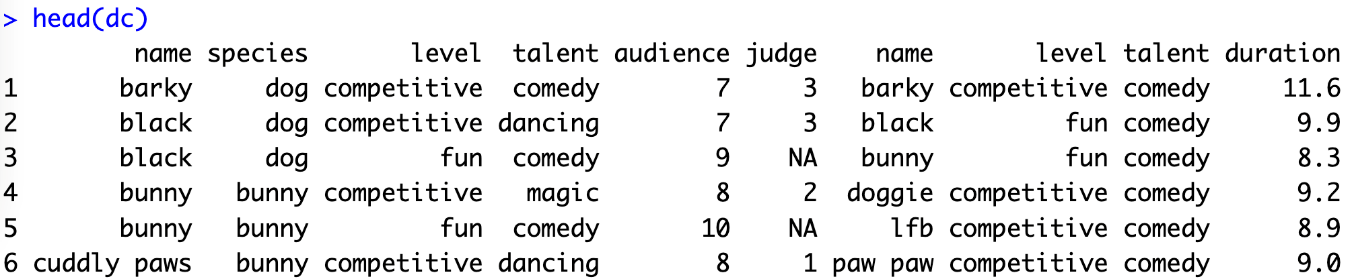
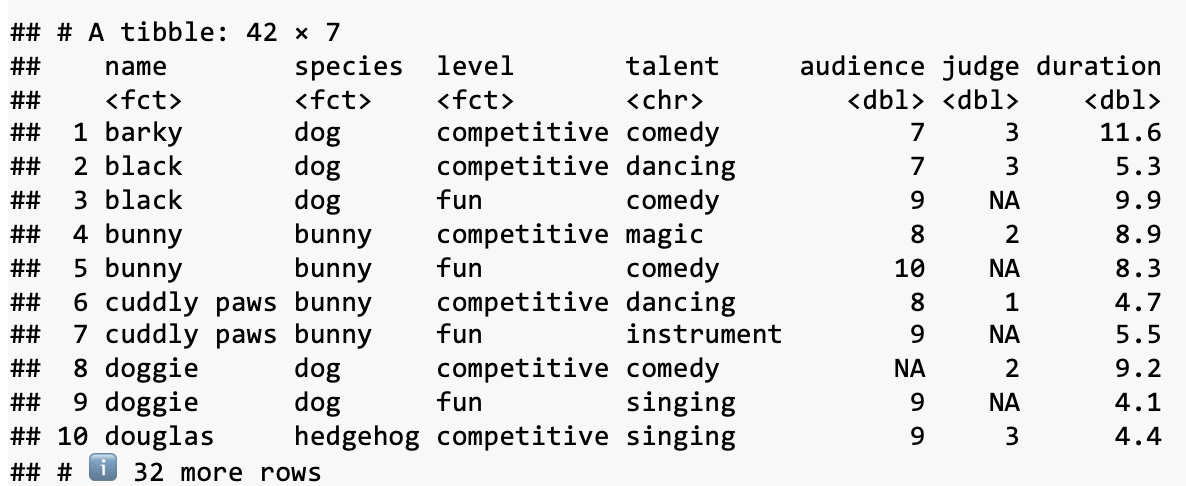


Combine d and dd together using the function full\_join(). We have not taught you this function so you will need to use your investigative skills to look it up and play around with it until you have figured it out. Assign the combined dataset to a new tibble called d\_full, and make it so the top rows of d\_full show up in the knitted Markdown. It should look like the screenshot below (rows may be in a different order, but the column order, column names2, size of the tibble, and data in each cell should be the same).

(c) The code given in the chunk here combines two tibbles by using the function cbind() rather than the function full\_join(). The output has been assigned to a tibble called dc whose output in the console is shown below. Based on a comparison of dc and d\_full, describe two major differences between what cbind() and full\_join() do, making clear reference to the parts of the tibbles that illustrate each difference. Finally, explain *why* these differences have occurred: how exactly cbind() combines tibbles that is different from how full\_join() combines tibbles.

*[Suggested word count: 90]*

2 Note that if you did not succeed in Q1(c) in renaming *kind* to *species,* your tibble here will have a column called *kind* instead. That is fine; you will only be penalised for this in Q1(c) and can still obtain full marks in Q4(b).

page6image57823088

Page 6 of 11

**Q5 [15% of total mark]**

(a) A tibble has been loaded for you called df, which is the same as d\_full. We are providing you with df here in case you weren’t able to create d\_full in Q4(b). Use the mutate() function along with case\_when() to make a new character variable in df called *durType.* [Note: We have not taught you case\_when()]. The value of *durType* is "long" if *duration* is more than 10, "short" if it is less than 5, and "medium" otherwise. Be sure to show the top of df in the knitted Markdown.

(b) Using only functions we have taught you, use df as the basis to create the tibble shown in the screenshot below. Assign it to the name ds, and make sure ds is visible in your knitted Markdown. Helpful hint: all of the variables are calculated from the *audience* variable. *medAud* indicates the median, and the others are self-explanatory.

(c) Based on the data in ds, **what talent is the least popular based on the mean audience ratings, and what is the least popular based on median audience ratings?** Why do the mean and median ratings for these give different results? Your answer should refer to the idea of central tendency that both mean and median each capture, and it should explain the discrepancy by relating this idea to the actual talent show data.

The mean audience rating show that magic (mnAud = 7.33) was the least popular talent, and for median audience rating it was instrument (medAud = 7.0).

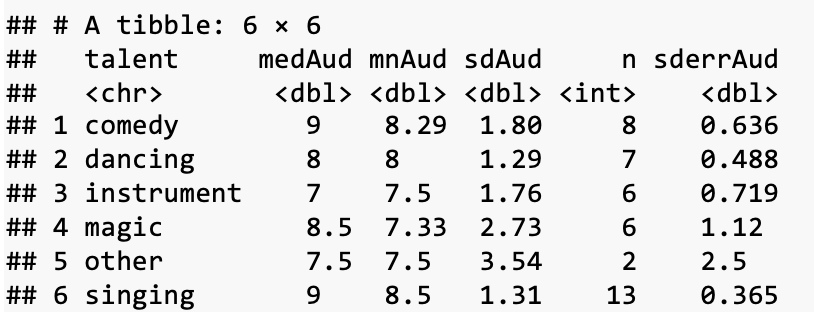
Mean and median use different approches for measuring typical tendencies in a set of data. Mean is found by adding all numbers together and then diving by number of observations, and median is the 50th percentile: the middle number when in ascending order.

The discrepancy in our dataset can be explained by the distribution of values: less variation for instrument, and more “extreme” scores for magic. Therefore, magic has both a higher medAud and lower mnAud than instrument.

The mean audience rating show that magic (mnAud = 7.33) was the least popular talent, and for median audience rating it was instrument (medAud = 7.0). Mean and median use different approaches for measuring typical tendencies in a set of data. Mean is found by adding all numbers together and then diving by number of observations, and median is the 50th percentule: the middle number when in ascending order. The discrepancy in our dataset can be explained by the distribution of values: less variation for instrument, and more "extreme" scores for magic. Therefore, magic has both a higher medAud and lower mnAud than instrument.

Based on the mean audience rating the least popular talent was magic, and for median audience rating it was instrument. The mean and median use different approaches for measuring typical tendencies in a set of data. The median is found by putting the values in ascending order, and then finding the value in the middle: in n=3, that would be nr. 2. And in n=4 we would find the mean of nr. 3 and 4. (e.g: (7+8)/2 = 7.5). On the other hand, we find the mean by addition of all values, divided by n. Mean is a better measurement tool for symmetrical data, and median is better for skewed data.

*[Suggested word count: 100]*



Page 7 of 11

**Q6 [12% of total mark]**

1. Make a bar plot like the one below using the d6 tibble, which has been loaded for you. For full credit, your figure should have all the components in the figure below (i.e., two panels, semi- transparent bars, dots, error bars, title, angled x-axis tick labels, three y-axis tick labels, etc.). Note that your individual data points will not be in exactly the same place as here because the geom introduces randomness; that is fine. The error bars should indicate one standard error. It’s fine if your colours aren’t exactly the same (you aren’t expected to guess what palette was used) as long as you use a sensible palette and theme, and the colours of the dots match the bars and vary as they do here. Note that if your knitted figure has a slightly different aspect ratio that is fine, as long as all of the elements are present and correct; different systems knit figures in slightly different ways.

d6\_barplot <- ggplot(data = d6, mapping = aes(x = talent, y = (medAud)), color = black, fill = talent, width = 0.7) +

geom\_bar(stat="identify", alpha = 0.5, size = 3, ) +

geom\_jitter(aes(color = talent), alpha = 0.6, size=3) +

geom\_errorbar(aes(ymin = medAud - sdAud, ymax = medAud + sdAud), width = 0.2, position = position\_dodge(0.5)) +

facet\_wrap(~level) +

theme\_bw() +

labs(subtitle = "Audience rating for each kind of talent",

x = "Talent",

y = "Rating (higher = better)") +

scale\_fill\_brewer(palette = set4)

print(d6\_barplot)

d6\_barplot <- ggplot(data = d6, aes(x = talent, y = medAud, fill = talent)) +

geom\_col(alpha = 0.5, color = "black", width = 0.7) +

geom\_jitter(aes(color = talent), position = position\_jitter(width = 0.1), size = 3, alpha = 0.6) +

geom\_errorbar(aes(ymin = medAud - sdAud, ymax = medAud + sdAud), width = 0.2, position = position\_dodge(0.7)) +

facet\_wrap(~level) +

theme\_bw() +

labs(

title = "Audience Ratings by Talent and Level",

subtitle = "Audience rating for each kind of talent",

x = "Talent",

y = "Rating (higher = better)"

) +

scale\_fill\_brewer(palette = "Pastel1")

print(d6\_barplot)

(b) Based on the graph in 6(a), describe any trends or regularities in performance that you observe. This is not a R question but rather a thought question asking you to critically think about what the data might be demonstrating and why this might be happening (you should speculate; just make sure to ground the speculation in the pattern of data and clearly indicate the part that is speculative). You’re not expected to make claims about significance but think about the *meaning* of the variables and discuss what (if anything) this figure might suggest about the talent show.

*[Suggested word count: 120]*

**Q7 [11% of total mark]**

1. Make a figure of your own using any of the tibbles provided (or any that you make from them if you want). Your goal is to show something new about the data that hasn't been shown by the previous figure. You should use at least one geom that you didn’t use in Q6, and you also need to incorporate two elements that you haven’t been taught in this subject. These can be anything from new geoms, a different palette package than RColor Brewer, a different theme, changing the size or style of your fonts, putting text inside the figure, changing aesthetic properties, or many other possibilities; you can do basically whatever you want as long as it’s new. The figure should have an informative title and axis labels, and a theme and colour palette other than the default. The aesthetic choices should add to its clarity rather than detract from it; part of what you are being marked on is if the figure illustrates the data in a clear and useful way.

df\_boxplot <- df %>%

ggplot(aes(

x = factor(judge),

y=duration),

fill = durType,

is.na(TRUE)) +

geom\_boxplot(aes(fill = durType), size = 3, outlier.shape = NULL) +

geom\_jitter(aes(color=durType), size = 2, alpha = 0.5, position = position\_jitter(width = 0.2)) +

facet\_grid(~level) +

theme\_bw() +

labs(

x = "Placement",

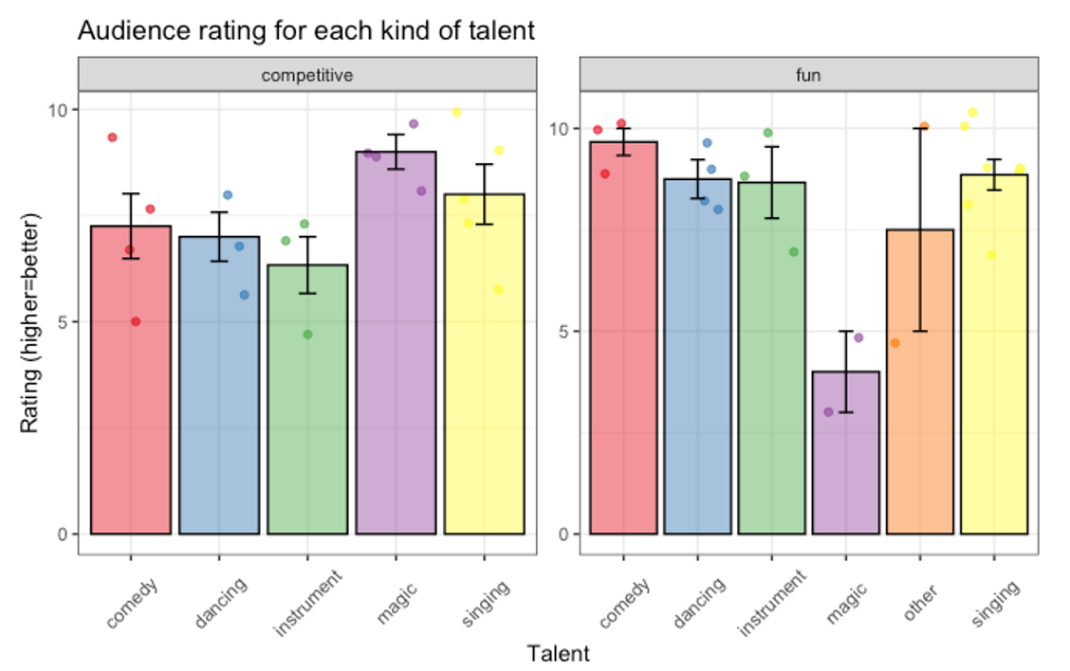
y = "Duration (minutes)",

title = "Duration and Judge-Placement") +

scale\_fill\_brewer(palette="Pastel2")

scale\_color\_brewer(palette="Pastel2")

print(df\_boxplot)



Page 8 of 11

(b) Explain what each of the two new elements are and how you made them. Your explanations don’t need to be extensive – for instance, if you hadn’t already been taught show.legend you might say “I got rid of the legend by adding show.legend=FALSE as an argument to the geom”.

*[Suggested word count: 50]*

(c) Explain what your figure suggests about the data. In your explanation be sure to describe the variables on each axis (and panel, if you have multiple panels) as well as what the pattern is and what it suggests about what is going on. (It is fine for you to say there is no pattern and it suggests that nothing much is happening if that is what you observe!) You won’t be evaluated on how interesting your result is, but on how clear and appropriate your explanation is given the figure. That said, it’s worth thinking about what kinds of research questions would be interesting to look at, since those are more likely to yield interesting patterns which are easier to discuss.

*[Suggested word count: 130]*

The filter-function returns rows that match our condition. This is a part of dplyr, which again is a part of the tidyverse-package. On the other hand, we have the base R-approach where the use of the logic functions display the values that match the criteria put forward through the logic function. We do not extract anything, but apply our criteria to the table. In both cases the data has to be 1 or 2 in the judge-column, and 8 or above in the audience-column to either be extracted or displayed. Filter does not return NA, but the logic function does.

**Q8 [3% of total mark]**

Gladly ran a statistical test and obtained a *p*-value of 0.07. “That means the null hypothesis is true according to the traditional alpha threshold of 0.05,” he explains. “However, I’m going to set my alpha threshold to be 0.1 instead; that will make the test statistic significant, so I can conclude the null hypothesis is false instead.” There are several distinct problems with Gladly’s idea. Explain two of them to him. For each, be sure to be clear about what the problem is and why it is a problem.

*[Suggested word count: 80]*

**Q9 [11% of total mark]**

You are provided with a code chunk that calculates the highest and lowest audience scores in our dataset (called *highest* and *lowest* respectively). Note also that part (b) and (c) use the tibble that you used in Q5 called df. Regardless of whether or not you succeeded in completing Q5, you can use df for Q9.

(a) Bunny observes that on average, in past talent shows about 70% of the audience sample has liked any given act. If we presume that average describes this talent show as well, what is the probability of observing the highest score we saw? The lowest? You should answer these questions using the function(s) taught in Week 5; you do not need to use any of the datasets themselves. Report probabilities as percentages, rounded to one decimal place.

(b) Gladly points out that they have other data from previous talent shows as well, not just about audience ratings. For instance, in previous years the average duration was 6.5 minutes, with a standard deviation of 3. Shadow, inspired, writes the code given to you in the code chunk. What does the calculated variable *prob* reflect? How is this related to the idea of a p-value? Is it possible to identify which individual data points are significantly different from previous averages? If so, which ones, and why? If not, why not?

*[Suggested word count: 100]*

(c) Can we draw conclusions about how significant the entire variable *duration* (i.e., the full dataset of data about *duration*) is, based on a single calculation combining only the individual *prob* values? If so, explain why. If not, explain why not and what other information is necessary. Note that you do not need to do any calculations here; this is a thought question about Week 5 concepts.

*[Suggested word count: 130]*

Page 9 of 11

**Q10 [8% of total mark]**

It’s evident from the data in Q6 that some kinds of talents have a much larger range of audience ratings than others. For instance, the range of magic tricks is 7 (i.e., with a low rating of 3 to a high rating of 10) while the range for singing and dancing is 4 (i.e., a low rating of 6 to a high of 10). Foxy starts wondering what kind of range one might *expect* to see in a random talent show, and how to determine if magic tricks are unusual.

Let’s help her out! Remember that one can have sampling distributions of any kind of statistic. We’ve spent a lot of time talking about the sampling distribution of the mean, but we could also think about the sampling distribution of the range, which applies when thinking about this question. In this problem you will reason about this situation, by direct analogy and extrapolation from what you’ve learned about the sampling distribution of the mean.

Foxy thinks that the *true* underlying distribution the audience ratings looks something like the figure directly below this paragraph: it’s very unlikely for 0 people to like a performance, slightly more likely for exactly 1 people to like it, and so forth, with it being most likely that 10 audience members like it. For the purposes of this question, let’s assume that she is correct and this is the true distribution.

1. Suppose talent shows become the next huge thing and as a result over the next few years there are 1000 talent shows. Each of the 1000 shows is divided into timeslots with 30 performances each. It is possible to calculate the range of audience rating for each of these timeslots.

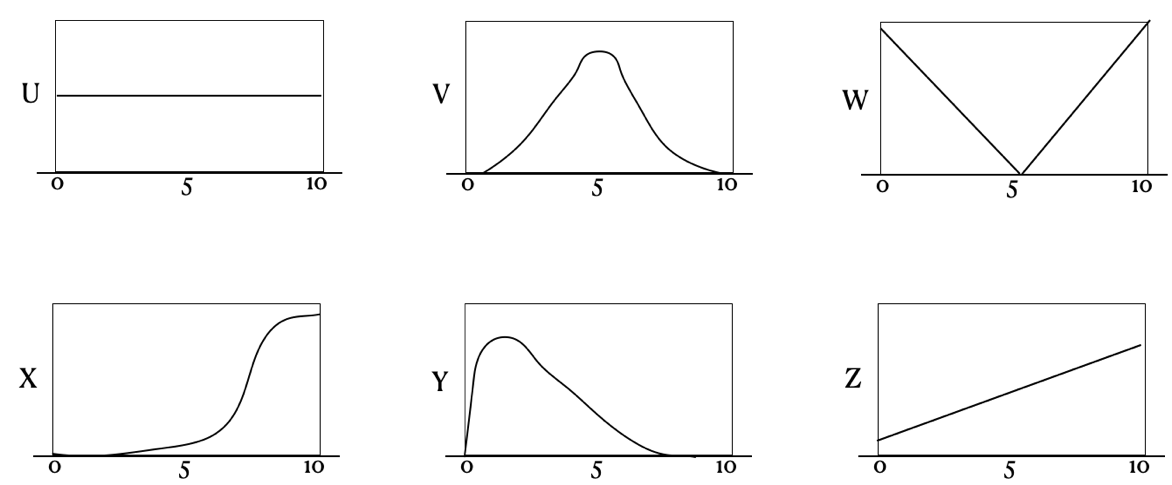
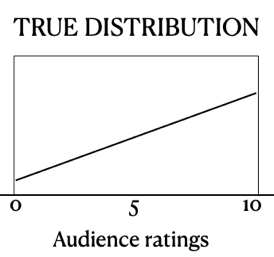
\*ANSWER: We have independent measures and a large enough sample (n < 30) the graph of sampling distribution of range (SDR) will move towards a normal distribution. We therefore expect the SDR to look like panel V, which shows a normal distribution. SDR == what we would expect the means of lots of samples from a population to look like. By using the timeslots we can find the SDR, which would be our estmiated population mean with less variance than our original distribution. [Word count: 91]\*

\*\*Q10b\*\*

\*ANSWER: This would not change my answer. Also here I expect the SDR to resemble panel V. The difference between sampling distribution of the mean and SDR is first of all that mean is a measure of central tendencies where as SDR is value reflecting the difference between the lowest and the highest values. Further,

Consider now the six panels U through Z below. Give the letter of the panel that most accurately captures what you expect the **sampling distribution of the range** to look like, on the assumption that the true distribution of audience ratings is as shown in the figure above. Explain your answer, making reference to the definition of sampling distribution and the figure. Hint: begin by thinking about what you would expect the range for a single timeslot of 30 performances to be.

*[Suggested word count: 100]*



Page 10 of 11

1. Suppose now that the underlying true distribution was uniform, as in the figure directly below this sentence.

With a true distribution being uniform, we would expect the SDR to resemble panel V. The difference between sampling distribution of the mean and SDR is first of all that mean is a measure of central tendencies where as SDR is a value reflecting the difference between the lowest and the highest values. A similarity between these two measures are that they may be susceptible to influence by outliers. With a uniform distribution all values would be as likely, which would make the values likely to group in the center (e.g. (1+2+5+8+9)/5 = 5) .

How would this change your answer to part (a), if at all? Considering the same panels U through Z, give the letter of the panel that you would pick as being the closest answer in this case. Explain why. How is the behaviour of the sampling distribution of the range similar to and different from the behaviour of the sampling distribution of the mean, as the shape of the underlying true distribution varies?

*[Suggested word count: 100]*

\* Note: You do not need to code or do any calculations in order to answer this question. This is a conceptual question designed to probe your knowledge about what a sampling distribution is. Moreover, if your intuition about the nature of a range are incorrect but your explanation of sampling distributions in general is solid, you can still get most of the partial credit.

**Q11 [2% of total mark]**

These marks are free as long as you say anything! What is your current theory about why everyone in Bunnyland is going hungry? (No word limit here, say as much or as little as you want)