


TPS REPORT 02

Organizing and Generating Data in Excel

1. Download the TPS 02 Excel file from www.terevaka.net/nau/ant568/tpso1.html
2. Navigate to the **Hour** sheet Suppose we have to interview prospective grad students at the next AAA meeting, and we are allowing each candidate exactly 5 minutes to impress us. On the **Hour** sheet, we are simply keeping track of the time each subsequent candidate begins her/his song and dance. In Cell **A1** enter 7:01 for the first time the first candidate begins. In Cell **B1**, enter 7:06 for the time the second candidate begins. Even the simple data entry in this process becomes pretty tedious if we must continue for 50 candidates. Excel has a handy feature for generating series-type data. First, select all Cells between **A1** and **AX1** to generate times for a total of 50 candidates. On the HOME tab at the top of your screen, use the  Fill ▾ button and select **Series...** to open a dialogue box. Notice that Excel automatically interprets the **Step value** for you (0.003472222222). That is actually converting your time difference between 7:06 and 7:01 to a decimal representing the fraction of a single day—5 minutes over 144,000 minutes (the number of minutes in a day).

$$0.003472222222 = \frac{5 \text{ min}}{(60 \text{ min/hr})(24 \text{ hr})}$$

Hit the OK button to complete the **Fill Series** action using the dialogue box. You should now have time listings for every five minutes between 7:01 and 11:06.

3. You can see on the **Candidate** sheet that we have already recorded all fifty candidates' names in the order in which they interviewed. Ideally we'd like to enter the time at which each candidate's interview begins next to her/his last name in Column **B**. Unfortunately, on the **Hour** sheet we just created that information in row format rather than column format. There are a few ways we could fix this. We could enter 7:01 and 7:06 in Cells **B2** and **B3**, respectively on the **Candidate** sheet and use the **Fill Series** command again. Instead, Copy Cells **A1** through **AX1** on the **Hour** sheet. Then right-click in Cell **A3** on the **Hour** sheet, and select **Paste Special...** and then the **Paste Special...** option to open a dialogue box. Check the **Transpose** box and then hit the OK button to complete the Paste. You should now have all your times transposed to column format. You could easily Copy those values and Paste them in Column **B** on the **Candidate** sheet to accomplish our goal. But there's another way to transpose data in Excel, using the **transpose** primitive, which is used in a special kind of formula in Excel called an *array formula*.

On the **Candidate** sheet, select all Cells between **B2** and **B51**. With that entire range still selected, immediately begin to type the array formula:


```
=transpose(Hour!A1:AX1)
```

The trick with array formulas is that you DO NOT simply hit enter to finish the formula. On a PC, you'll use Ctrl+Shift+Enter (on a Mac, you'll use Command+Shift+Return).

If you used the **transpose** primitive, you should now have values in Cells **B2** through **B51**, but they might be back in that funky decimal format. If so, select these same cells, right-click on the highlighted range, and select **Format Cells...** to bring up a dialogue box. You'll want to convert all these decimals to the **Time** Category (specifically the **Type 13:30** option). Then hit the OK button to complete the **Format Cells...** action.

Now add a heading in Cell **B1**.

3. Interviewing 50 candidates and comparing their accomplishments takes a long time, and there are quicker ways we could select the next round of grad students. Suppose we want to make decisions based on the length of candidates' last names. In Cell **C2**, write a formula that will calculate the length of the last name in Cell **A2**.
 - Write your formula in your TPS Report.
4. Now Copy your formula from Cell **C2** and Paste to all Cells between **C3** and **C51**. In Cell **C53**, write a formula to calculate the maximum number of characters for a last name in the entire pool of candidates.
 - Write your formula in your TPS Report.
5. Your calculated value in Cell **C53** should be 9. Unfortunately, that doesn't identify *which* candidate has the longest last name. This becomes an even bigger problem for much larger sets of data.

Select Cells **C2** through **C51**, and then hit the  Conditional Formatting dropdown button to select the **Highlight Cells Rule...** and then the **Equal To...** option. **Format cells that are EQUAL TO: 9** with Green Fill with Dark Green Text. Hit the Ok button to complete the **Conditional Formatting** action.


Now it should be much easier in your data to identify the candidate with the longest last name of all.

Add a heading in Cell **C1**.

6. Now we'll learn how to generate data randomly in Excel. Add the heading "GPA" in Cell **D1**. Let's suppose our candidates' GPAs range between 0.0 and 4.0. We could

use the **randbetween** primitive in a formula that will calculate a random integer between 0 and 4. But realistically, GPAs usually come in a numeric format with two decimal places. So, instead of calculating an integer, write a formula in Cell **D2** that will calculate a random rational number between 0.00 and 4.00, always providing output with two decimal places.

➤ Write your formula in your TPS Report.

7. Copy your formula from Cell **D2** and Paste to Cells **D2** through **D51**. We have to be careful with some formulas using primitives like **randbetween** in Excel because they can update or recalculate random cell values at unexpected times. Try deleting the contents of Cell **D2** and retyping your formula there. You'll see that when you hit Enter, all of your randomly generated numbers in that column will be recalculated—not just the value of Cell **D2**. One way to avoid automatic updates is to Copy all of Column **D**, then right-click on the highlighted region, and **Paste Values**.
8. There's another way to generate random numbers in Excel. Delete the contents of Cells **D2** through **D51**. Now go to the  **Data Analysis** button on the **DATA** tab at the top of your screen (you must have the Data Analysis Toolpak installed). Select **Random Number Generation** from the **Analysis Tools** and hit the OK button.

In the dialogue box, we want to generate 1 **Variable** (GPA), but we want to generate 50 **Random Numbers**.

We want all values between 0.00 and 4.00 to be equally likely so we'll select **Uniform** for our **Distribution**.

Our **Parameters** should be set to **Between 0 and 4**.

The **Random Seed** can be used to generate totally random numbers, but to generate the same sequence of random numbers every time a certain seed is used. We'll leave that blank for now.

Excel gives us options of where we want our calculated random numbers to appear (in a range, a new sheet, or even a new file). Select **Output Range** and let's use the range **\$D\$2:\$D\$51**. Then hit the OK button to complete the **Random Number Generation** action.

9. Now we have a new problem. Our random numbers have more than two decimal places. Write a formula using the **round** primitive in Cell **E2**, to convert the random number in Cell **D2** to a number with only two decimal places.

➤ Write your formula in your TPS Report.

Now Copy your formula from Cell **E2** and Paste to Cells **E3** through **E51**.

10. Now Copy all of Column E, right-click on the highlighted area, and **Paste Values**.
11. Now Copy all of Column D, right-click on the highlighted area and **Delete**.

Add a new “GPA” header to Cell **D1**.

12. Write a formula using the **rank** primitive in Cell **E2** to calculate the GPA rank for candidate Smith in the entire pool of GPAs.

➤ Write your formula in your TPS Report.

Now Copy your formula and Paste it to Cells **E2** through **E51** (Hint: Be careful to use the \$ sign before the row numbers in your **ref** argument within the **rank** formula. Otherwise your pool of GPAs will effectively change for each cell/row you Paste into!!!).

13. In Excel, we can refer directly to the contents of any cell from any other cell. For example, if we want Cell **H2** to calculate the contents of Cell **A5**, we could simply enter in **H2**:

=A5

Which, in this case would result in the string “Brown”. We can also refer to the contents of any other cell in a more roundabout way using the **index** and **indirect** primitives. For example, we could just as easily replace:

=A2

With:

=indirect(left(A13,1)&len(A2))

In other words, that would be like referring to the Cell with the alphanumeric label with the single, leftmost character of the contents of Cell **A13** and the number of characters in the contents of Cell **A2**. Or in this case, that formula entered into Cell **H2** would result in the string “Brown”. The **index** primitive is similar in that it calculates the contents of another cell indirectly. For example, if we wanted to pick a random last name from the 50 last names listed in the contents of Cells **A2** through **A51**, we could enter in Cell **I2**:

=index(A\$2:A\$51,randbetween(1,50))

If we wanted to repeat this random selection process from our list of last names 9 more times, we could Copy this formula and Paste it to all Cells between **I3** and **I11**.

14. But what if we want to make sure we never get the same name twice when picking a random last name, or we want to *sample* the list of last names *without replacement*? For example if we gave up on making any real effort to rate candidates, and just

decided we were going to admit ten random candidates, the process wouldn't make any sense if we were able to pick the same last name twice (or sample *with replacement*). Thinking up an Excel formula to sample our list of last names without replacement is considerably harder than the last formula. But it is possible for you now that you know the **index** and **rank** primitives.


To do this, first use the primitive **rand** to generate a random number between 0 and 1 for each cell between Cell **G2** and **G51**. Notice that the:

```
=rand( )
```

formula accepts no arguments inside the parentheses. Copy that formula and Paste it to Cells **G2** through **G51**. Each one of those random numbers is totally unique, and so each cell between **G2** and **G51** would also have a unique rank between 1 and 50. So if we consider the ranks of those cells, we have essentially just randomized the order of all integers between 1 and 50. So if we want to create a formula to sample our list of last names 10 times without replacement, we could write the formula:

```
=index(A$2:A$51,rank(G2,G$2:G$51))
```

In Cell **J2** and then Copy it and Paste it in Cells **J3** through **J11**.

15. Make sure that each one of your columns has a header in it, then make sure all of your headers are in **bold** font.
16. Select all Cells between **A1** and **J51**. Create bright red borders around and between all of those cells using the  dropdown button and selecting **More Borders...**

Now Copy that table with borders and

- Paste it into your TPS report so that it fits on the same page.

{ *transpose, randbetween, rank, round, rand, index, indirect* }