Seattle Children's®
HOSPITAL • RESEARCH • FOUNDATION

# Do not use means on Likert scale data

**Seattle Children's Hospital**
***Analytic Guidelines Series***

**Dwight Barry, PhD**
*Enterprise Analytics*

*17 December 2016*

## Summary

- Likert and similar ordinal-level scales have a variety of uses, particularly within surveys such as Family Experience, Culture of Safety, and Employee Engagement. They also occur in clinical care, for example, in the use of pain scores.
- When evaluated improperly—particularly through the use of averages—the results can be strikingly misleading. Obviously, misleading results could drive or promote action where none is warranted, and vice versa.
- In nearly all cases, not only is it mathematically wrong, **taking the average of a Likert-scale variable will *not* provide useful answers** to the questions end-users can use to make actionable decisions. In essence, the use of averages cannot account for the importance of capturing and understanding variabililty. Analysts should strive to avoid their use in any reporting solution or analytic product that uses ordinal-scale data.
- Better ways to represent ordinal-value results include histograms of the values themselves, the use of well-supported "top-box"-type proportions, and/or bar charts of percentage by score or score category (e.g., favorable/neutral/unfavorable).
- "Statistical significance" on changes or differences between response groups' medians or distribution shift can be assessed through non-parametric frequentist tests (permutation, Mann-Whitney-Wilcoxon), Information Theory, or Bayesian analysis. $t$-tests should never be used on Likert scales because ordinal data does not meet the assumptions of a $t$-test. Also, when using frequentist tools, one must *also* account for multiple testing to reduce the chance of false positives.
- A good way to remember not to use means on Likert scale data is to think: "The average of Good and Excellent is *not* Good-And-A-Half".

# Discussion

*Note: all of the data in this document is fake, created using random number generators specifically to illustrate particular points.*

## A simple example

Take a simple example where a group of 6 people people take the same survey for 4 years, and the mean results for an important question, such as "my team works well together", are as follows:

Taking the mean of these results gives you this:

| Year 1 | Year 2 | Year 3 | Year 4 |
|---|---|---|---|
| 4.1666667 | 4.3333333 | 4.3333333 | 4.3333333 |

So one might conclude that there is an improvement from year 1 to year 2, and no change year-over-year after year 2.

The values that created the above results are as follows:

| Individual | Year 1 | Year 2 | Year 3 | Year 4 |
|---|---|---|---|---|
| A | 5 | 2 | 3 | 1 |
| B | 4 | 5 | 5 | 5 |
| C | 4 | 5 | 5 | 5 |
| D | 4 | 5 | 5 | 5 |
| E | 4 | 5 | 5 | 5 |
| F | 4 | 4 | 3 | 5 |

You might already see how management decisions would be made differently based on whether one had just the means or had the complete data.

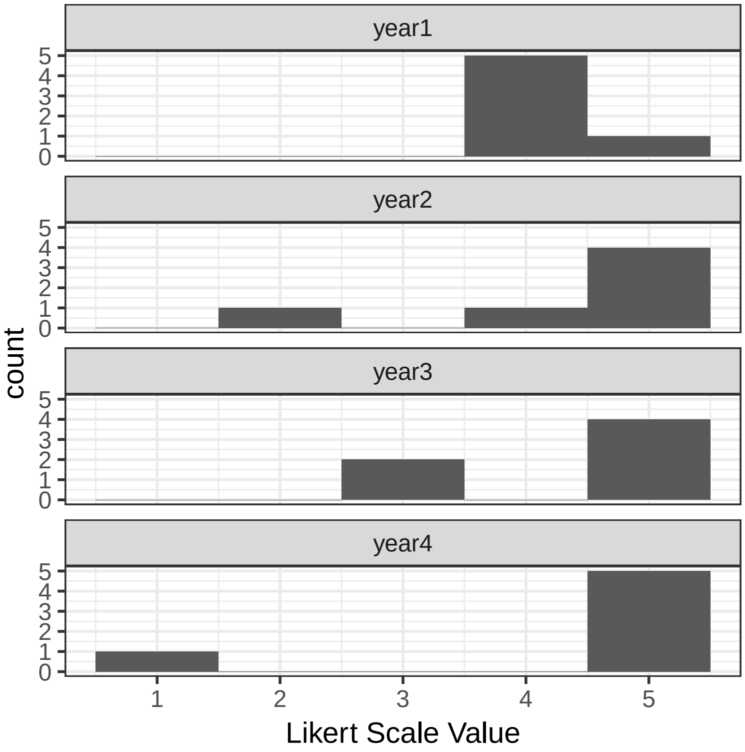
However, in the latter case, you risk reducing or eliminating anonymity, which is essential toward getting respondents to answer truthfully (not to mention being unethical). Further, poring over tables of answers for many people for long surveys makes that approach practically infeasible. Visualizing the results in ways that capture a more complete story provides an answer to both issues, as well as providing decision-makers with truly-actionable information.

## Visualizing Likert-scale data

### Histograms

Histograms of the actual score values are the best way to visualize Likert data—they have two real axes, showing counts by score value or category, so you can parse the visual and understand the results very quickly. Using the same data as above, one can instantly see that the "improvement" in year 2 was perhaps not an improvement after all; while most respondents appear to be satisfied above what they thought in year 1, one respondent may be at risk of leaving.

*Figure 1. Histogram of example Likert scale data.*



**Likert charts**

The main disadvantage of histograms is space; Likert charts—which are in essence just stacked bar charts—are far more compact. The disadvantage is that it takes slightly longer for a user to parse them, but when faced with lots of questions or groupings, they tend to be the best option.

There are two kinds of Likert charts—those that use a center line for a point of reference, and those that do not, in which case they are simply percentage bar charts or mosaic plots. In the graphs below, each score value has its own color, and each score category—e.g., unfavorable is 1-2, neutral is 3, and favorable is 4-5 on a 5-point scale—is summarized by a percentage value at the left, middle/interior, and right sides of the bar, respectively.

*Figure 2. Centered Likert chart.*

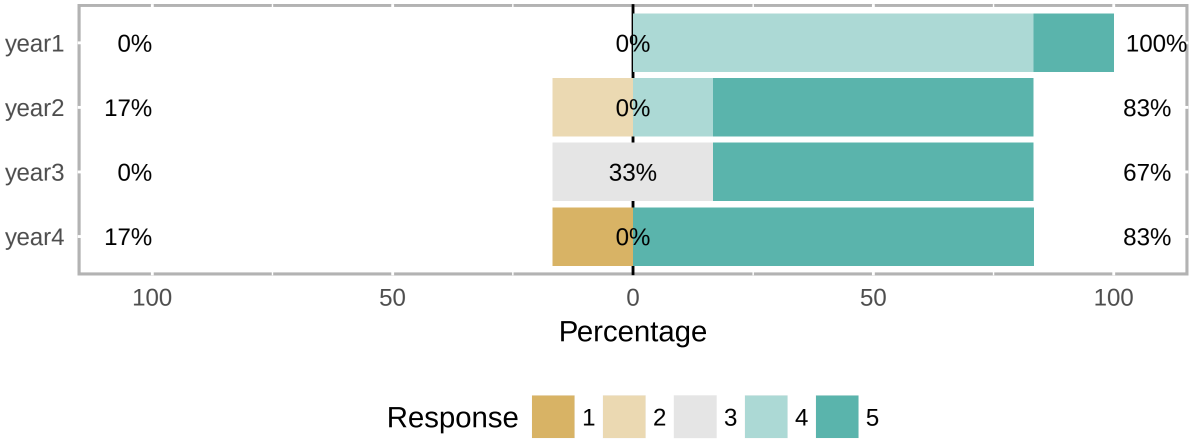
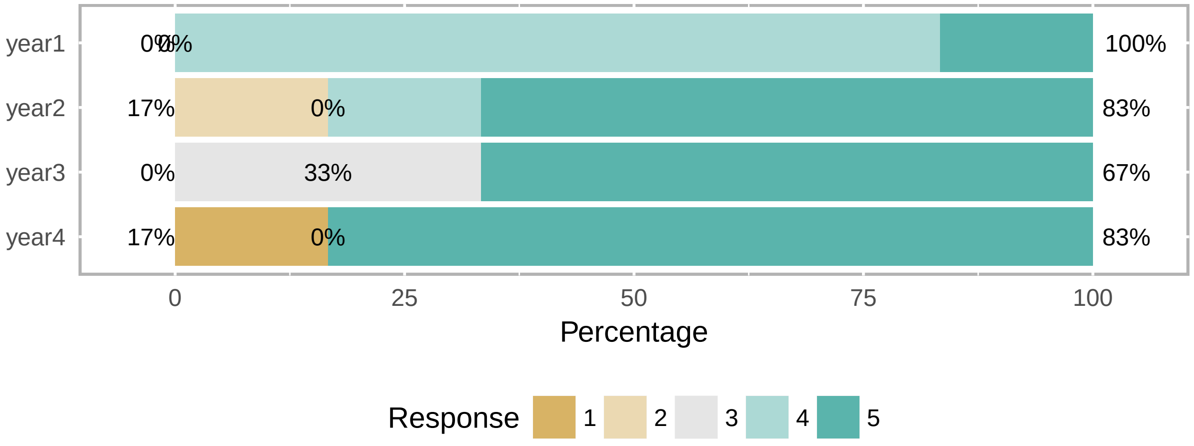


*Figure 3. Uncentered Likert chart (aka percent bar chart).*



Neither Likert chart type does as good a job as the histogram at making the results immediately understandable, but again, histograms take more space, and busy decision makers often need to see the forest (all the questions) at the expense of some trees (each question). In this case, analysts might use the histograms to explore potentially important results themselves, and then use Likert charts in a report with some strategically-placed text highlighting important patterns they found with the histograms.

## How many respondents are enough?

It's common to think: "We surveyed everyone in this department, therefore the results we see must be correct." However, how people responded to surveys depends on many factors—such as mood the date the survey is taken, recent events in life and in work, changes in organizational structure, and any number of other factors—and many internal surveys are given only once a year. Thus, survey results are really a *sample* of attitudes and opinions, subject to random events and natural fluctuations.

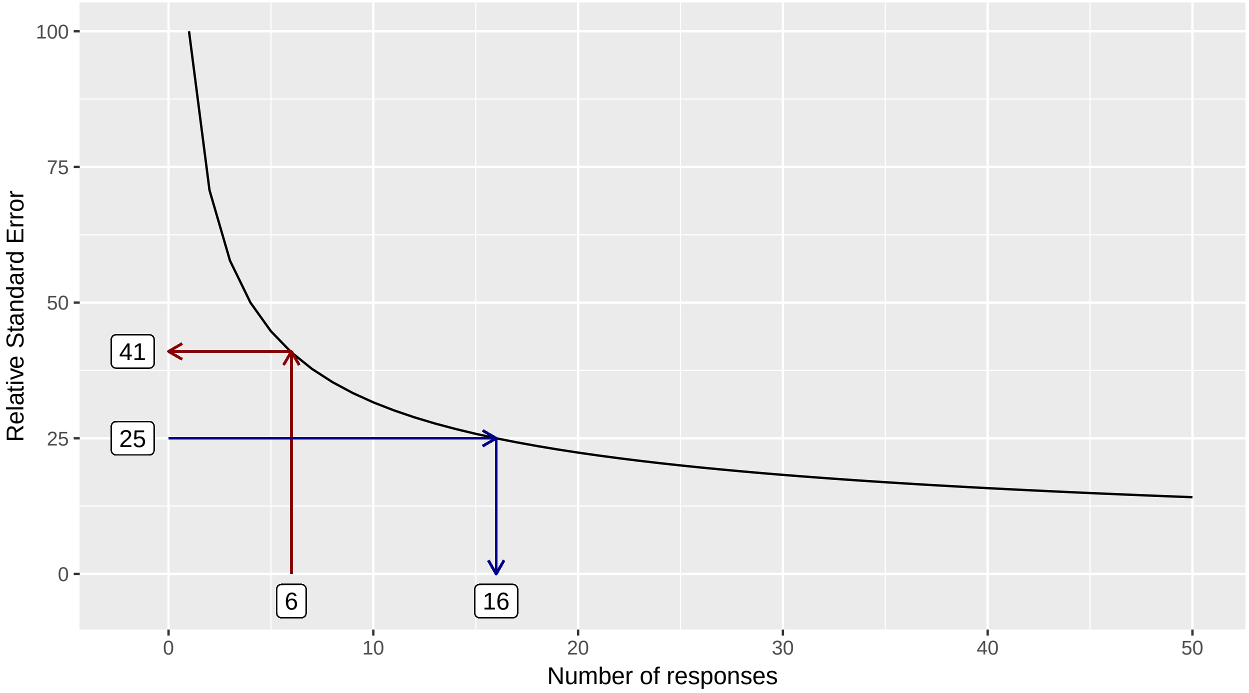
Typical practice at SCH is to expose summary results for groups with at least 6 people. While this helps preserve some anonymity, it does not include enough responses to ensure the overall response is stable. Comparisons over time or across groups that are not based on stable results can lead to incorrect conclusions about differences that may or may not be real.

In this context, *stable* means that the data accurately represent true changes (or lack of change) in the question at hand. It's basically impossible to distinguish natural variation from real change when you have small numbers of respondents. As a result, the National Center for Health Statistics, for example, does not publish results with less than 20 distinct cases or responses.

The relative standard error (RSE) is the metric used to evaluate whether you have enough values for the results to be stable. The standard error is an estimate of the likely difference between the results and the true value (which in surveys, even of complete populations, can't be known exactly due to the reasons mentioned above). The *relative* standard error is the standard error expressed as a percent of the measure or number of responses, which is a constant function: $\frac{1}{\sqrt{responses}} * 100$. This function can be seen in the graph on the next page.

Generally, you want RSE values less than 20-25% to have some confidence that your results are stable.

*Figure 4. The RSE-response count function. The RSE associated with the use of 6 responses is marked with dark red, and the response count associated with an RSE of 25% is marked with dark blue.*



## Is there a significant difference?

Many decision-makers want to know if a result is "statistically-significantly different" from, say, the same response from a previous time period, or between a couple of subgroups in the same response. Unfortunately, this is mostly useless, for two reasons.

First, acting as if Likert or other ordinal scales are continuous level data leads to many problems of interpretation (see the Appendix for a general overview of measurement scales and appropriate statistics). There has been controversy over this distinction for many decades; however, a great way to understand the conceptual problem is to realize that the mean of *Agree* and *Strongly Agree* is **not** *Agree-And-A-Half*—it just makes no sense.

A subsequent argument might be that, no, it's not conceptually accurate, but it provides a sense for directional changes. However, such results still run into problems of interpretation: if you go from 4.16 to 4.33, have you gone from Agree.16 to Agree.33? What does such an "improvement" mean, in practical terms? All you can accurately say is that both values are most consistent with an *Agree* opinion.

Specifically in the medicine/healthcare context, Kuzon et al. state that the use of parametric statistics on ordinal data (such calculating a mean or using a *t*-test) is the first of "The seven deadly sins of statistical analysis". Don't "sin" and you don't have to worry about whether your results are illegitmate.

One way around this is to use medians and test for differences in those statistics (with medians, the difference is best assessed via bootstrap or permutation testing), to test whether the distribution has shifted (Mann-Whitney-Wilcoxon test), or to use more advanced techniques such as multinomial or proportional-odds regression (see the Advanced section, below). These options are the more statistically-correct ways to do it, as opposed a *t*-test.

So, using the simple example above, we'd want to know whether the median is statistically different between year 1 (Median = 4) and year 2 (Median = 5). Running a permutation test gives us the following results:

```
>
>   Exact Two-Sample Fisher-Pitman Permutation Test
>
>  data:  value by variable (year1, year2)
>  Z = -0.33333, p-value = 1
>  alternative hypothesis: true mu is not equal to 0
```

While our effect size is “1”—more accurately, *Agree* to *Strongly Agree*—the *p*-value of the test is very large (basically 1), so we cannot say that this difference is “statistically significant”.

We could also ask, “has the distribution shifted?”, which would involve using the Mann-Whitney-Wilcoxon test:

```
>
>   Wilcoxon rank sum test with continuity correction
>
>  data:  value by variable
>  W = 11.5, p-value = 0.285
>  alternative hypothesis: true location shift is not equal to 0
```

The *p*-value is again non-significant, so the change between year 1 and year 2 can’t be assumed to be a statistically significant change. Looking at the raw data or visuals, a decision-maker might be justified in wanting to act, but the analysis suggests that the difference is not statistically significant.

This leads us to the second problem with using *p*-values for determining whether a statistically-significant difference has occurred: sample size.

*p*-values are directly dependent on sample size. If your sample is large enough, you are guaranteed to have a low *p*-value. If your sample is small, whether or not you get a significant *p*-value depends on the scale of difference between the groups, i.e., the effect size.

For example, consider the following examples evaluating the number of people who answer *Agree* or *Strongly Agree* (the “favorable” score group) to a question:

| Example | Favorable | Total Answers | Effect size | *p*-value |
|---|---:|---:|:---:|:---:|
| 1 | 15 | 20 | 75% | 0.04 |
| 2 | 114 | 200 | 57% | 0.04 |
| 3 | 1,046 | 2,000 | 52% | 0.04 |
| 4 | 1,001,450 | 2,000,000 | 50% | 0.04 |

With 15 of 20 people selecting a favorable value on the Likert scale, we have an effect size of 75%, which is an effect worth taking seriously. That value is also a statistically significant difference ($p < 0.05$), which supports the idea that the majority has a favorable opinion. With a couple of thousand responses (example 3), we again have a statistically significant difference, but the effect size is now only 52%, close enough to even-preference as to be *practically* the same. In medical terms, we might think of this as statistically significant but clinically irrelevant.

For these—and many other reasons outside the scope of these guidelines—statisticians are moving away from the use of *p*-values. In frequentist statistics, these are being replaced by the use of effect sizes and confidence intervals (CIs); these provide information on both on the precision of the estimated difference, as well as whether the difference can be considered statistically distinct. If the CI includes 0, the difference is not-significant. Regardless of the location of 0, the width of the CI tells you how precise your estimate is.

```
>  Difference in medians is 1.

>  BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
>  Based on 1000 bootstrap replicates
>
>  CALL :
>  boot.ci(boot.out = median_diff, type = "perc")
>
>  Intervals :
>  Level     Percentile
>  95%   (-1,  1 )
>  Calculations and Intervals on Original Scale
```

Here, we see that the effect size is a difference in medians of 1, but the confidence interval on that effect size goes from -1 to +1, i.e. is consistent with any score difference between *Neutral* and *Strongly Agree*. Since that CI includes 0, we can't say that the change from median of *Agree* to a median of *Strongly Agree* is statistically different, though again, sample size matters—one would probably like to try to intervene based on the one respondent who dropped down to 2 (*Disagree*) anyway.

## ***Neutral*** scores matter

You might have noticed in some surveys that there is often no longer a "neutral" or "undecided" category included in the middle of the scale, e.g., what's usually a 3 on a 5-point Likert scale. Sometimes it is placed at the end of the scale, and sometimes it is eliminated entirely. The reason for this is that those terms can sometimes be interpreted in a variety of ways; for example, with a question such as "My pay is fair compared with other companies", a *Neutral* response could indicate "I'm neutral on this", "yes, I guess so", "I don't know", "it's neither fair nor unfair", "I don't want to answer", "I'm not sure what 'fair' means", and any number of ideas that don't necessarily indicate a true neutral opinion.

When a question has a response option where this type of ambiguity exists, a mean value will tend toward the that option because of this bias, unless of course the mean is already at that value. However, when *Neutral* is marked as 3, and when valid responses tend towards 4s and 5s, these ambiguous responses will drag down the average (and vice versa for responses heavy with 1s and 2s). Of course, you shouldn't use means anyway, as we've seen above, but many reports do—so understanding this effect is important toward interpreting the results in a useful way.

Use of a median is somewhat resistant to this problem, though you still won't know whether the middle values are valid responses or accidents of interpretation.

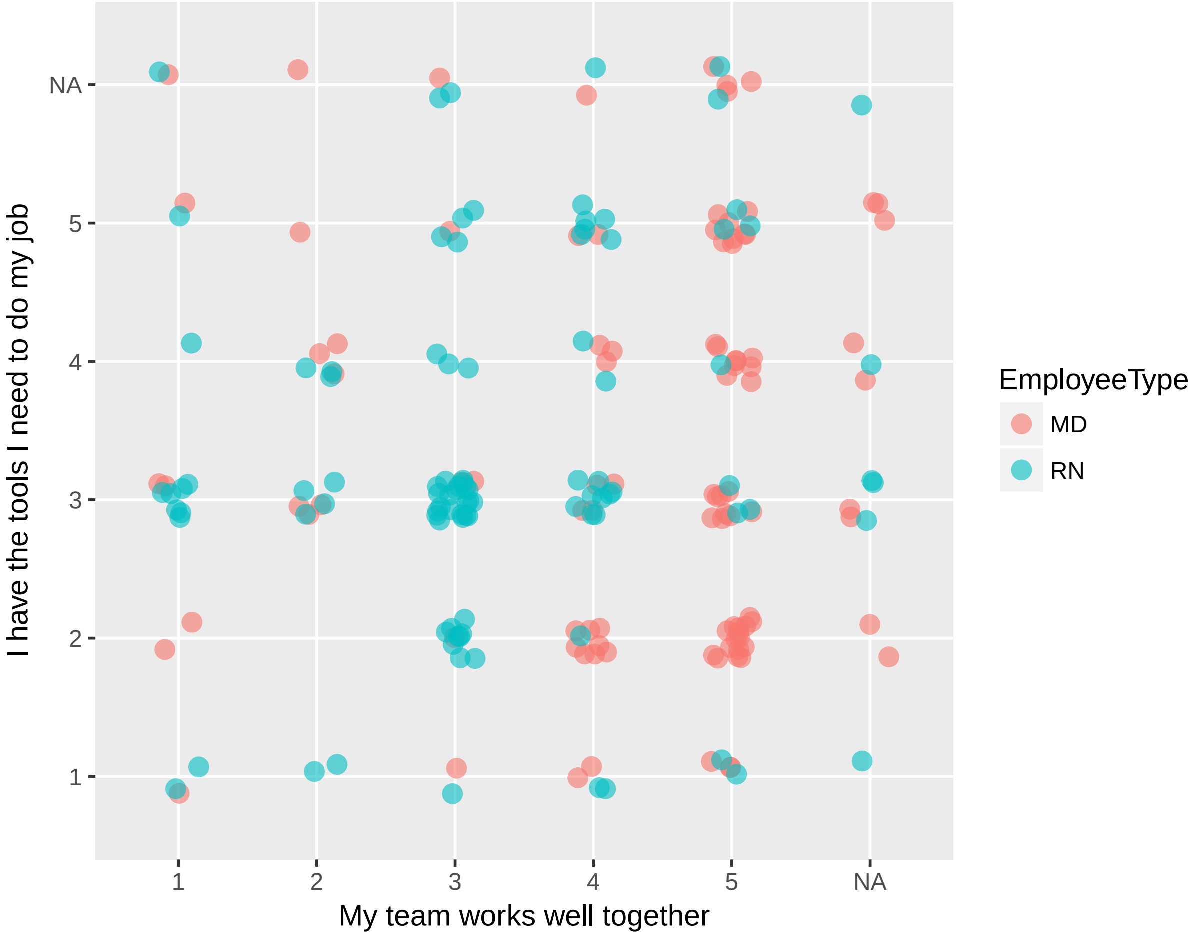
When you see an "undecided" or "N/A" response placed at the end of the scale, it is usually (but not always!) a sign that the survey creator understands this problem.

Sometimes, of course, *Neutral* can be a completely reasonable and unambiguous response to a question. Context matters; while it's easiest for survey creators and scanning software to use the same scale for large numbers of questions, it is important that the analyst understands the extent to which *Neutral* and similar types of responses are a valid part of the measurement scale.

## Similarities: correlation between ordinal-scale variables

Although traditionally many analysts used non-parametric correlation like Spearman's or Kendall's, polychoric correlation is the proper tool to assess similarities between Likert scale results. (Polyserial correlation is used when one variable is numeric and the other is ordinal.)

*Figure 5. Scatterplot of ordinal comparisons (jittered to show point density) between the questions "My team works well together" and "I have the tools I need to do my job".*



The polychoric correlation coefficient between "My team works well together" and "I have the tools I need to do my job" is 0.0579. As expected, that suggests that there is no relationship between the responses to these two questions.

## Other ordinal-scale visualizations

*Figure 6. A Likert chart for two different questions (e.g., as within a single year's survey), with a count histogram to show number of responses and non-answers for each question.*

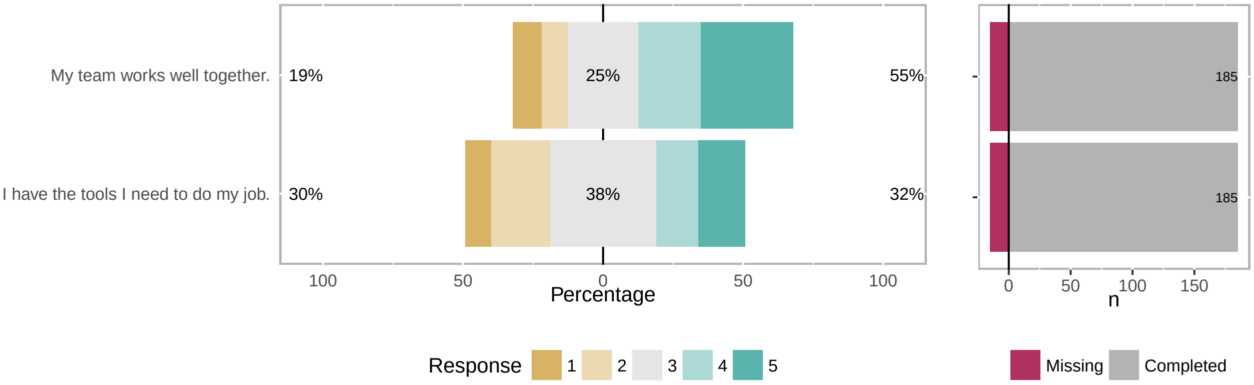


*Figure 7. An uncentered Likert chart for two different questions.*

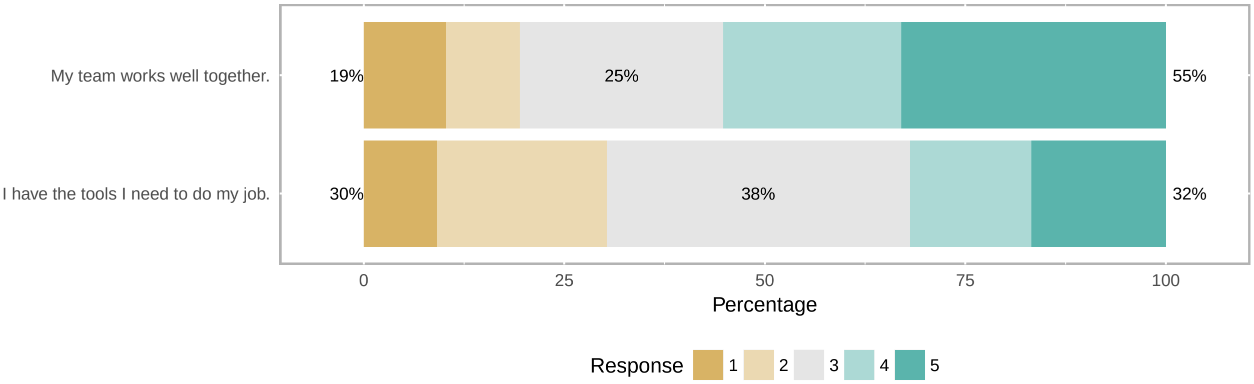


*Figure 8. A heatmap of the response frequency for two different questions. While the use of means and SDs is inappropriate, this particular example directly illustrates why those values don't capture the response patterns in the data.*

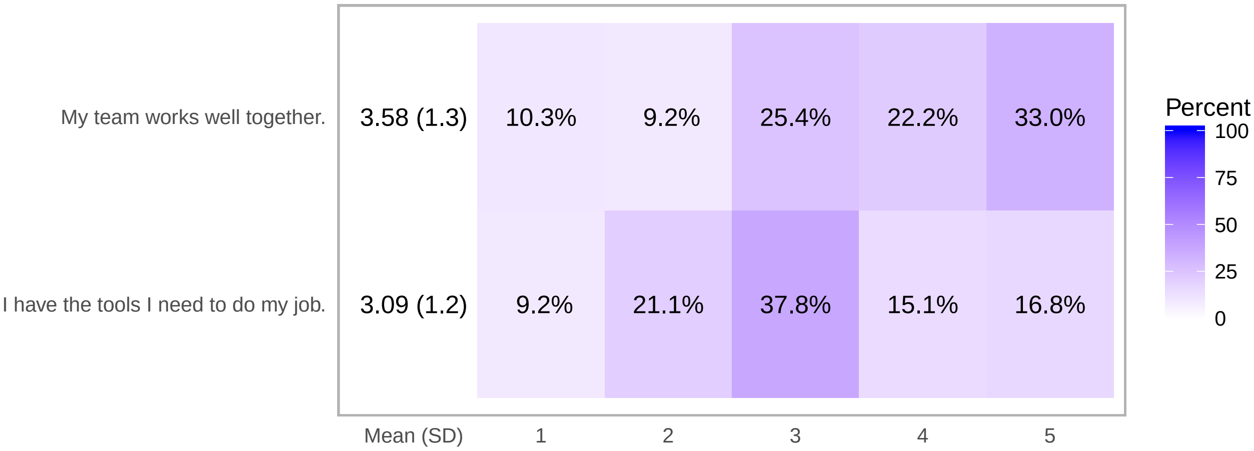


*Figure 9.* *Subgroups can sometimes reveal patterns not seen in aggregate data. For example, compare the overall results for "My team works well together" in Figure 5 (above) with the responses from the subgroups of MDs and RNs (below, bottom panel).*

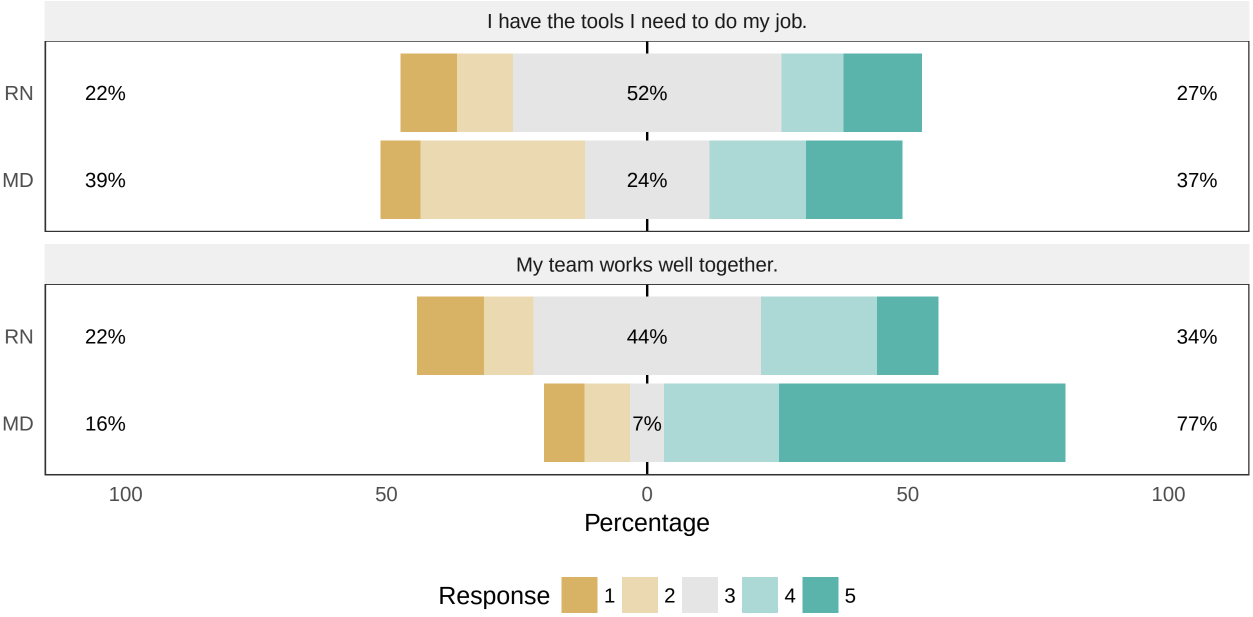
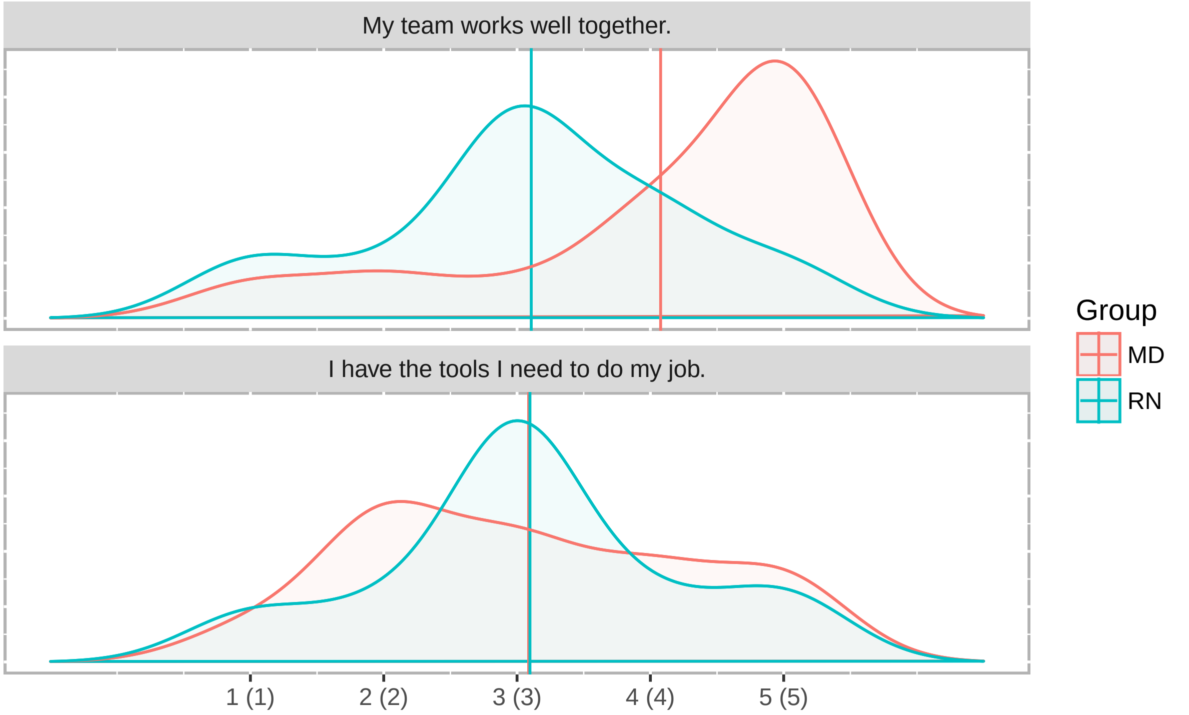


*Figure 10.* *Density plots for the same data shown in Figure 8, above. While using a density plot on ordinal data is also statistically inappropriate, it can be a useful tool for an analyst. Bar histograms are difficult to overlay subgroups or different years for a direct comparsion, so must be separated into facets instead (e.g., Figure 1, above). Density plots are easier to overlay to show these comparisons, so while not appropriate for a report, they can be useful tools for an analyst during the exploration phase.*

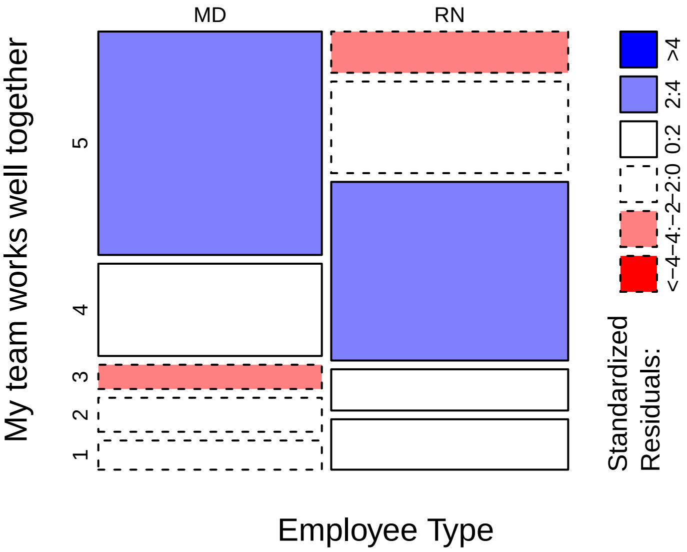


## Advanced analytics

While Mann-Whitney-Wilcoxon (sometimes known as the Mann-Whitney $U$-test) is the test most often used with differences between ordinal distributions, there are other options that can tell you whether a measured difference between groups is statistical different.

The old stand-by in this case is the $\chi^2$ test, which is often best visualized with a mosaic plot.

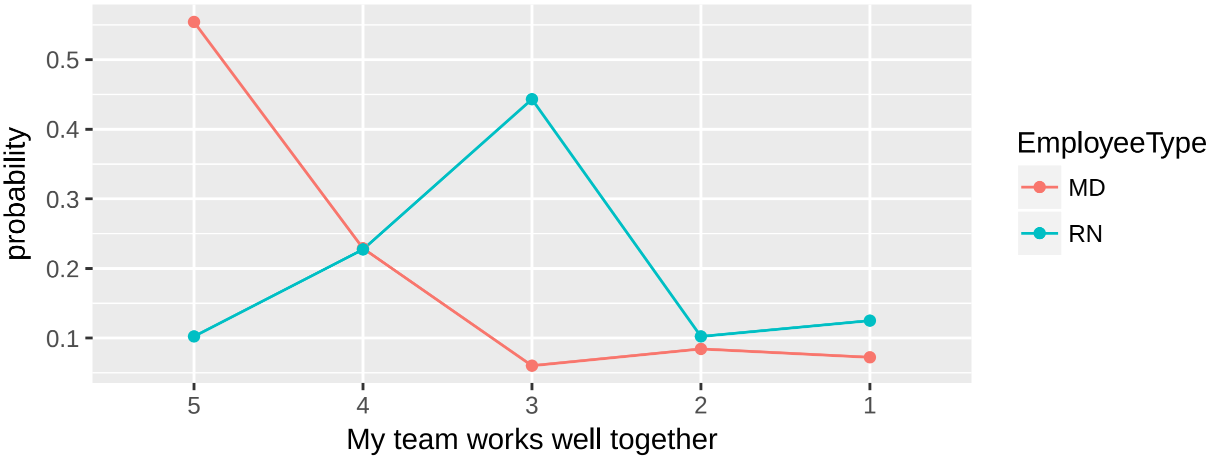
*Figure 11. Chi-square test and mosaic plot between Employee Type and responses to the "My team works well together" question.*



```
>
>   Pearson's Chi-squared test with simulated p-value (based on 2000
>   replicates)
>
>  data:  both2_tab
>  X-squared = 52.809, df = NA, p-value = 0.0004998
```

The multinomial regression model is a more powerful (and more modern) version of the $\chi^2$ test.

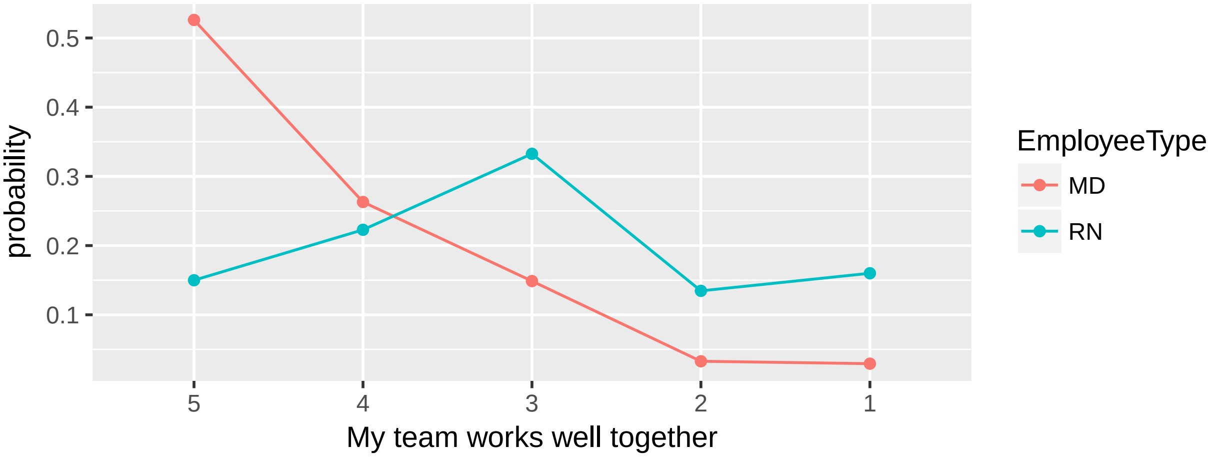
*Figure 12. Multinomial regression between Employee Type and responses to the "My team works well together" question, with information-theoretic table for multi-model inference.*



| Modnames | K | AICc | Delta_AICc | ModelLik | AICcWt | LL | Cum.Wt |
|---|---|---|---|---|---|---|---|
| Employee Type | 8 | 472.0249 | 0.00000 | 1 | 1 | -227.5680 | 1 |
| Null Model | 4 | 522.0647 | 50.03976 | 0 | 0 | -256.9118 | 1 |

If you can meet the assumptions, the proportional-odds regression is more powerful than the multinomial model, as it can take into account the ordered nature of the ordinal scale.

*Figure 13. Proportional odds logistic regression between Employee Type and responses to the "My team works well together" question, with information-theoretic table for multi-model inference.*



| Modnames | K | AICc | Delta_AICc | ModelLik | AICcWt | LL | Cum.Wt |
|---|---|---|---|---|---|---|---|
| Employee Type | 5 | 485.9008 | 0.00000 | 1 | 1 | -237.7686 | 1 |
| Null Model | 4 | 522.0647 | 36.16387 | 0 | 0 | -256.9118 | 1 |

If the concepts or ideas in this section are confusing, it's probably worth consulting a statistician for help evaluating your data with these tools.

## Appendix: Measurement Levels & Appropriate Summary Statistics

| Statistic / Parameter | Categorical *Nominal* | Ranked *Ordinal* | Discrete/Counts *Interval/Ratio* | Continuous *Interval/Ratio* |
|---|---|---|---|---|
| Data set size (n) | Y | Y | Y | Y |
| Percent / Frequency | Y | Y | Y | Y |
| Count or rate | Y | Y | Y | Y |
| Categories (levels) | Y | Y | Y | Y |
| Mode | Y | Y | Y | Y |
| Median | *No* | Y | Y | Y |
| Interquartile range | *No* | Y | Y | Y |
| Median absolute deviation | *No* | Y | Y | Y |
| Range | *No* | Y | Y | Y |
| Minimum/maximum value | *No* | Y | Y | Y |
| Quantiles | *No* | Y | Y | Y |
| Mean (average) | *No* | *No* | Y | Y |
| Standard deviation | *No* | *No* | Y* | Y* |
| Coefficient of variation | *No* | *No* | Y* | Y* |

* You must use the correct distribution (proper mean-variance relationship) to ensure you get the correct standard deviation; most software defaults to calculating the standard deviation for a normally-distributed sample, which could be incorrect for certain kinds of count, rate, or proportion data, for example.