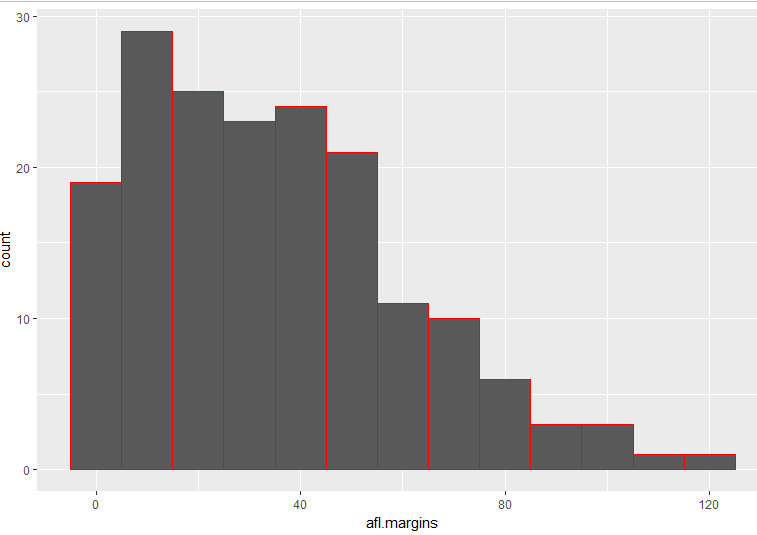
***Learning Statistics with R - University of Adelaide***

***Part IV - Working with Data***

**Chapter 5 - Descriptive Stats**

**Chapter 5.1 – Measures of Center**

* Any time you get a new data set to look at, 1 of the first tasks to do is find ways of summarizing the data in a compact, easily-understood fashion.
* This is what **descriptive statistics** (as opposed to **inferential statistics**) is all about.
* In fact, to many people, the term "statistics" is synonymous w/ descriptive statistics.



* This histogram is **right/positively-skewed** (*tail to the right*)
* This indicates that higher margins are less frequent = use median as center of measure/measure of central tendency b/c w/ a skewed distribution, the mean is pulled to the side of the tail (to the right)
* **Mean** = center of mass/gravity of a data set (balancing point if histogram was on a seesaw)
* If data are **nominal**, *don’t* use mean OR median, since both rely on the idea that *the numbers assigned to values are meaningful*
* If the numbering scheme is *arbitrary*, it is probably best to use **mode** instead.
* If data are **ordinal,** we’re more likely to want to use **median**, which only makes use of the ordering info within data (i.e., which numbers are bigger), but does not depend on the *precise* numbers involved
* **Mean**, on the other hand, makes use of the *precise numeric values* assigned to observations, so it is not really appropriate for ordinal data.
* For **interval** and **ratio** data, either mean or median is generally acceptable, depending on what you are trying to achieve.
* mean advantage = it uses ALL the info in the data (useful when you don’t have a lot of data)
* mean disadvantage = *very sensitive* to extreme values (**outliers**)
* There’re systematic differences between mean + median when a distribution is *asymmetric/skewed*.
* Median is closer to the "body" of the distribution/histogram, whereas mean is dragged towards the tail (extreme values)
* 1 of the fundamental rules of applied statistics = **the data are messy**, since real life is never simple, so data sets you obtain are never as straightforward as the statistical theory says.
* **Robust statistics** = tries to grapple w/ the messiness of real data + develop theory to cope w/ it.
* Ex: 100, 2, 3, 4, 5, 6, 7, 8, 9, 10 --> If observed in real life data set, probably suspect something funny is going on w/ the “100”, it’s probably an outlier
* Consider removing it from the data set
* In real life, however, you don't always get such cut-and-dried examples.
* For instance, we might get (15, 2, 3, 4, 5, 6, 7, 8, 9, 12), and 15 may look suspicious, but nowhere as much as 100
* It’s little trickier, as it might be a legitimate observation, it might not.
* When faced with a situation where some of the most extreme-valued observations might not be quite trustworthy, the mean is not necessarily a good measure of central tendency
* b/c it’s highly sensitive to even just 1 or 2 extreme values, + is thus NOT considered to be a **robust measure**
* 1 remedy to this 🡺 *Use the median*
* For more general solution = use a **trimmed mean** -🡪 "discard" the most extreme examples on both ends (largest + smallest values) of the distribution and take the mean what remains.
* Goal of trimmed mean = preserve best characteristics of the mean + median
* It’s not highly influenced by extreme outliers, but like the mean, it uses more than 1 observation
* We generally describe a trimmed mean in terms of the % of observations on either side that’re discarded.
* Ex: **10% trimmed mean** discards largest + smallest 10% of observations + takes mean of the remaining 80%
* 0% trimmed mean = regular mean, 50% trimmed mean = median.
* In this sense, a **trimmed means** = provides a whole family of central tendency measures that span the range from regular mean to the median
* Fairly substantial difference between mean + median = the mean may be influenced a bit too much by extreme values
* **Mode** of dataset = value of the variable that makes up the mode
* **Modal frequency =** actual number of occurrences of the mode.'
* Mode is most often calculated for **nominal** data (means + medians are useless for those variables)
* Plus, there are some situations in which you really DO want to know the mode of an ordinal, such as **interval** or **ratio** scale variable.
* Ex: afl.margins values clearly indicate a **ratio** (i.e. point differential field), so in most situations mean or median is the measure of central tendency we want.
* But consider a friend offering a bet to pick a football game at random + you have to guess the exact margin.
* If you guess correctly, win $50, if not, lose $1
* For this bet, mean + median are completely useless, + it is the mode you should bet on

**Chapter 5.2 - Measures of Variability**

* Statistics discussed so far all relate to **central tendency**, which all talk about which values are "in the middle" or "popular" in the data.
* Central tendency is NOT the only type of summary statistic to calculate,
* 2nd thing we want = a **measure of variability** of data, or how *spread out* data are, how *f*ar away from the mean/median do observed values tend to be?
* Although range = *simplest* way to quantify variability, it's one of the *worst* b/c it’s not robust
* If the data has 1 or 2 extremely bad values in it, we would like our stats not to be unduly influenced by these cases, but unfortunately range is
* Ex: (100, 2, 3, 4, 5, 6, 7, 8, 9, 10) 🡺 range = 110 + *is not robust*, but if the outlier 100 were removed, we would have a range of only 8.
* **Interquartile Range (IQR)** is like the range, but instead difference between biggest + smallest value, its difference between 25th quartile/percentiles (Q1) + 75th quartile/percentiles (Q3).
* Reminder: *10th percentile of a data set = smallest number, x, such that 10% of data is less than x*
* So, median = 50th quartile/percentile (2nd quartile, Q2)
* It’s obvious how to interpret the range, a little less obvious to interpret IQR.
* Simplest way to think about it: **IQR** = range spanned by the *middle half* of the data
* 1/4 of the data falls both below + above the 25th + 75th percentiles, leaving the "middle half" of the data between the 2.
* The 2 measures so far, range + IQR, both rely on the idea that we can measure the spread of the data by looking at the quartiles of the data.
* *However, this isn't the only way to think about the problem.*
* A different approach is to select a meaningful **reference point** (usually mean or median) + then *report the "typical" deviations from that reference point.*
* A typical deviation = usually the mean or median value *of these deviations*
* This leads to 2 different measures = **mean absolute deviation (AAD)** + **median absolute deviation (MAD)**



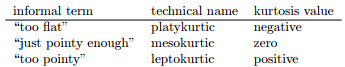
* **MAD** seems to be used in statistics + does seem to be the better of the 2
* **AAD** does occasionally show up in psychology though.
* Although **AAD** has its uses, it's NOT the best measure of variability to use.
* From a purely *mathematical* perspective, there are some solid reasons to prefer *squared* deviations rather than absolute deviations
* This obtains a measure, **variance, s^2**, which has a lot of really nice statistical properties but 1 massive psychological flaw
* It’s basically same formula as **AAD**, except w/ *squares of deviations* rather than *absolute values*
* **NOTE**: Variances are **additive**
* X and Y have variances VarX and VarY
* Define a new variable Z that = sum of X + Y
* It turns out **VarZ = VarX + VarY** is a very useful property (but not true of other measures)
* There is a subtle distinction between describing a sample + making guesses about a population from which the sample came
* Up to this point, it's been a distinction without a difference.
* Regardless of whether describing a sample or drawing inferences about a population, *mean is calculated the same*
* NOT SO for variance or standard deviation, or for many other measures
* Most of the time we’re not interested in the sample in + of itself but we have a sample exist to tell you something about the world.
* In other words, we’re starting to move away from a \*\***statistic**\*\* to a \*\***parameter**\*\*.
* How do you interpret the variance?
* Descriptive statistics are supposed to describe things + right now the variance is a gibberish number.
* *There really is no human-friendly interpretation of variance, which is the most serious problem w/ it*
* Although it has some elegant mathematical properties that suggest it is a fundamental quantity for expressing variation, it is completely useless if you want to communicate w/ an actual human.
* Variances are completely uninterpretable *in terms of the original variable*
* If all the numbers have been squared, they do not mean anything anymore, which is a huge issue.
* Ex: Game 1 Margin was 376.36 points-squared higher than the average margin
* There is not a real unit of measurement
* Solution: Take the *square root of the variance* = \*\***standard deviation**\*\* = \*\***root mean squared deviation (RMSD)**\*\*
* "A standard deviation of 18.01 points" = much easier to understand b/c expressed in original units
* Interpreting SDs = slightly more complex b/c SD is derived from variance, a quantity w/ little to no meaning to humans
* As a consequence, we rely on a simple rule of thumb:
* In general, expect 68% of data to fall w/in 1 SD of the mean, 95% to fall within 2 SD, and 99.7% to fall within 3 SD
* Tends to work well most of the time, but it's not exact + is calculated *based on an assumption that the data is symmetric/normal/bell-shaped* (AFL winning margins is not)
* 65.3% of AFL margins lies w/in 1 SD, still pretty consistent w/ the "approximately 68%” rule
* **Median absolute deviation (MAD)** =pretty much identical to the idea behind the mean absolute deviation (AAD) but w/ the median
* Has a straightforward interpretation*: Every observation in the data lies some distance away from the typical value* (here, the median).
* MAD = an attempt to *describe a typical deviation from a typical value in the data set.*
* It wouldn't be unreasonable to interpret MAD = 19.5 for AFL margins by saying:
* *"The median winning margin in 2010 was 30.5, indicating a typical game involved a winning margin of about 30 points. However, there was a fair amount of variation from game to game, w/ a MAD value = 19.5, indicating a typical winning margin would differ from this median value by about 19-20 points.'*
* Although a "raw" MAD value is completely interpretable on its own terms, it is not actually how it's used in a lot of real world contexts.
* Instead, a researcher *actually* wants to calculate the *SD*, but since the mean is very sensitive to extreme values, so too is the SD
* So, just like median = "robust" way of calculating "something like the mean", it's not uncommon to use MAD = "something that is like the SD".
* Unfortunately, a raw MAD value of 19.5 doesn't do this when if our SD = 26.07.
* However under certain assumptions, we can *multiply the raw MAD value by 1.4826* + obtain a # directly comparable to the SD
* As a consequence, a default value of this constant = 1.4826
* **NOTE:** If you want to use this "*corrected*" MAD value as a robust version of the SD, you really are relying on the assumption that the data is normal (which is not true for afl.margins)
* Summary\*\*
* **Range** = full spread of data, very vulnerable to outliers, + as a consequence is not often used unless you have good reasons to care about extremes in the data.
* **IQR** = where the "middle half" of the data sits, is pretty robust, complements median nicely
* **Mean absolute deviation (AAD)** = how far observations are on average from the mean, very interpretable, but has a few minor issues (not discussed here) that make it less attractive to statisticians than the SD, so it is used sometimes, but not often.
* **Variance** = average squared deviation from the mean, mathematically elegant, is probably the "right" way to describe variation around a mean, but is completely uninterpretable b/c it does not use the same units as the data, so almost never used except as a mathematical tool but is buried "under the hood" of a very large number of statistical tools.
* **SD** = square root of the variance, fairly elegant mathematically, expressed in same units as data, so is interpreted pretty well, in situations where mean is the measure of central tendency, is the default + by far the most popular measure of variation.
* **Median absolute deviation (MAD)** = typical (i.e. median) deviation from the median value, is simple + interpretable in *raw* form, is a robust way to estimate SD in the *corrected* form, used for *\*some\** kinds of data sets, not used very often, but does get reported sometimes.
* In short, the **IQR** + the **SD** are easily the 2 most common measures used to report variability of data, but there’re situations in which the others are used.

**Chapter 5.3 – Skew and Kurtosis**

* In practice, skew nor kurtosis is used as frequently as measures of central tendency + variability
* **Skew** = pretty important, mentioned a fair bit; but **kurtosis =** rarely reported in a scientific manner
* **Skewness** = a measure of asymmetry
* Data w/ a lot of extreme small values 🡪 lower tail = "longer" than upper w/ few extremely large values = **negatively/left skewed.**
* More extremely large values than extremely small ones = **positively/right skewed**.
* Actual formula for the skewness of a data set:



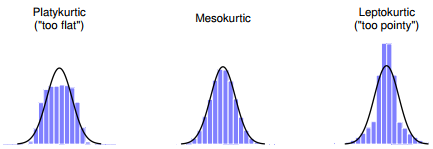
* Not surprisingly, AFL winning margins data is fairly skewed, as we saw before in the histogram
* **Kurtosis** = a measure of the "**pointiness**" of a data set
* By convention, **normal curve** = 0 kurtosis = **mesokurtic** (just pointy enough)
* So pointiness of a data set is assessed *relative to the normal curve.*
* Not pointy enough (bit of a uniform distribution) = negative kurtosis = **platykurtic**
* Too pointy (large peak relative to all else) = kurtosis = positive = **leptokurtic**



* The equation for kurtosis is pretty similar in spirit to formulas seen already for variance + skewness:

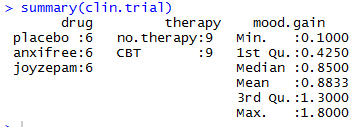


* Afl.margins is close to **mesokurtic** = just pointy enough (0.2962)



**Chapter 5.4 – Summarizing a Dataset**

* Pass a data frame to summary() = produces slightly condensed summary of each variable inside

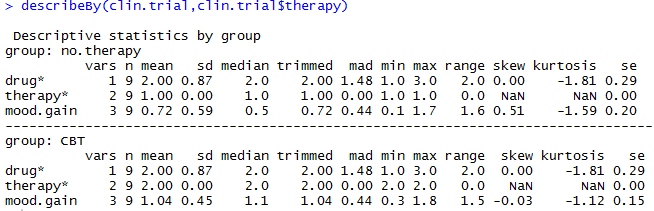


* Describe() output:
* **n =** sample size (more precisely, the number of non-missing values)
* **mean =** *sample* mean
* **sd** = the (*bias-corrected*) standard deviation
* **median**
* **trimmed** = 10% trimmed mean (default)
* **mad =** median absolute deviation
* **min, max, range**
* **skew, kurtosis**
* **se** = standard error of the mean
* **describe**() (in psych package) = only useful when data = interval or ratio scale (*encoded as numeric vectors*)
* For nominal or ordinal variables (**factor** vectors), these descriptive statistics are not useful
* describe() converts factors + logical variables to numeric vectors in order to do the calculations.
* Such variables are marked w/ \*’s and most of the time, the descriptive statistics for those variables won’t make much sense.
* If you try to feed it a data frame that includes a character vector as a variable, it produces an error.

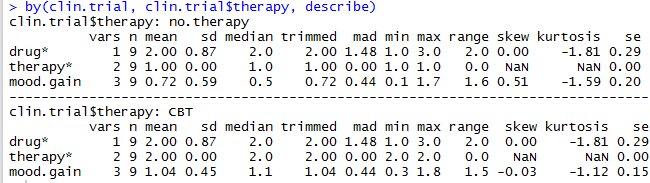


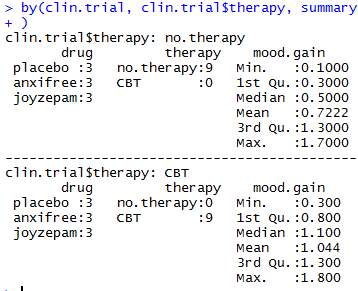
**Chapter 5.5 – Descriptive Statistics Separately For Each Group**

* Look at descriptive statistics, broken down by some grouping variable

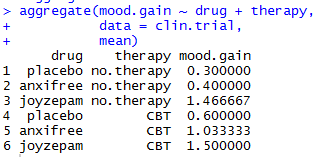
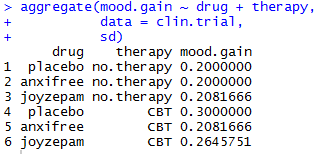


* See 2 outputs, 1 for the CBT group + 1 for the “no therapy” group.
* output still displays asterisks for factor variables
* Somewhat more general solution is offered by the by() function





* What if you have multiple grouping variables?
* Suppose, for example, you would like to look at average mood gain separately for all possible combinations of drug + therapy.
* Possible w/ by() and describeBy() functions used in conjunction, but it’s usually more convenient to use aggregate()

**Chapter 5.6 - Standard Scores**

* Ex: Questionnaire intended to measure “grumpiness” w/ 50 questions, which you can answer in a grumpy way or not.
* Across a big sample (hypothetically 1M), the data are fairly normally distributed, w/ mean = 17/50 questions answered in a grumpy way + SD = 5.
* In contrast, I answer 35/50 questions in a grumpy way. So, how grumpy am I?
* Could say I have *grumpiness* of 35/50, so you might say I’m 70% grumpy
* If my friend had phrased questions a bit differently, people might have answered in a different way, so the overall distribution of answers could easily move up/down depending on the precise way in which questions were asked.
* So, I’m only 70% grumpy with respect to *this set of survey questions.*
* Even if it’s a very good questionnaire, this isn’t a very informative statement.
* A simpler way = describe MY grumpiness *by comparing me to other people*.
* *“Out of the sample of 1M people, only 159 people were as grumpy as me, suggesting I’m in the top 0.016% of people for grumpiness. “*
* This makes much more sense than trying to interpret raw data.
* This idea that we should describe *my* grumpiness in terms of the *overall distribution of the grumpiness of humans* = the qualitative idea that **standardization** attempts to get at.
* 1 way to do this is to *describe everything in terms of* ***percentiles***.
* However, the problem w/ doing this is that “it’s lonely at the top”.
* Suppose that my friend had only collected a sample of 1k people (still a pretty big sample for the purposes of testing a new questionnaire) + this time got a mean = 16/50 w/ SD = 5
* The problem is that almost certainly, not a single person in that sample would be as grumpy as me.
* Different approach = convert my grumpiness score into a **standard score/z-score**
* **Standard z-score** = *the number of SDs above/below the mean that a score lies.*





* We can now transform a raw grumpiness into a standardized grumpiness score



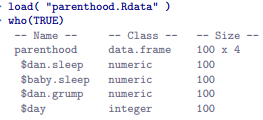
* To interpret this, recall the rough heuristic 99.7% of values are expected to lie w/in 3 SDs of the mean, so if my grumpiness corresponds to a z score = 3.6 🡪 indicates I’m very grumpy indeed
* **pnorm()** allows us to be a bit more + calculates a **theoretical percentile rank** for a score based on a z-score

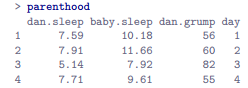


* This suggests that I’m grumpier than 99.98% of people.
* In addition to allowing you to interpret a raw score in relation to a larger population (+ thereby allowing you to make sense of variables that lie on arbitrary scales), standard scores serve a 2nd useful function 🡪 can be *compared to one another in situations where raw scores can’t.*
* Suppose *another* questionnaire measured extraversion using a 24-item questionnaire.
* The overall mean for this measure = 13 w/ SD 4 + I scored a 2
* Doesn’t make a lot of sense to try to compare a raw score of 2 for extraversion to a raw score of 35 for grumpiness.
* Raw scores for the 2 variables are about fundamentally different things = apples to oranges.
* z for grumpiness = 3.6 + z for extraversion = -2.75 🡪These 2 #’s CAN be compared to each other
* *Much less* extraverted than most people (z = -2.75) + much grumpier than most people (z = 3.6)
* But the *extent* of my unusualness is much more extreme for grumpiness (3.6 > 2.75).
* B/c each standardized score is a statement about where an observation falls *relative to its own population*, it is possible to compare standardized scores across completely different variables.

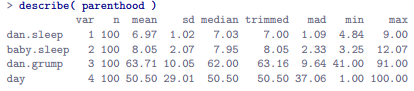
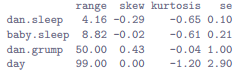
**Chapter 5.7 – Correlations**

* *Relationships* between variables in data = the **correlation** between variables.
* When calculating means + SDs from a *sample* of data, but actually want to talk about a score *relative to a population*, what I’m actually doing is estimating a z score.
* Though some caution is usually warranted.
* It’s not always the case that 1 SD on variable A corresponds to the same “kind” of thing as 1 SD on variable B.
* Use common sense when trying to determine whether or not z scores of 2 variables can be meaningfully compared.
* Suppose I’m curious to find out how much my infant son’s sleeping habits affect my mood.
* I can rate my grumpiness very precisely, on a scale from 0-100 + also assume I’ve been measuring my grumpiness, sleeping patterns + my son’s sleeping patterns for 100 days
* Let’s say, for 100 days, I’ve saved the data as a file called parenthood.Rdata.



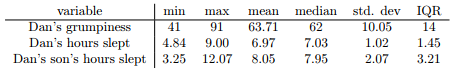


* Next, calculate some basic descriptive statistics:

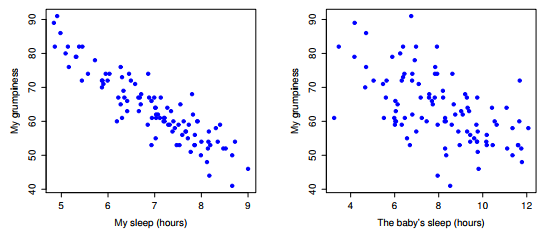
 



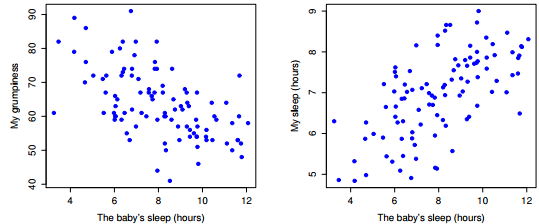
* ***NOTE:*** Just b/c you can calculate dozens of different statistics doesn’t mean to report all of them
* If writing up for a report, pick out statistics that are of most interest to readership + put them into a nice, simple table



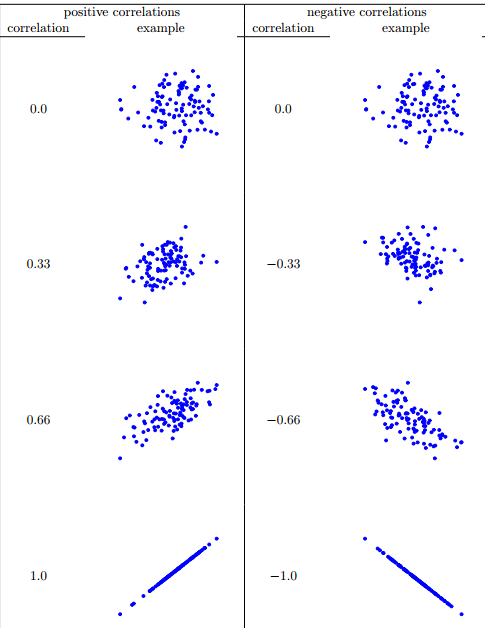
* For a table = every variable has “human-readable” names, always good practice
* Even this table is more than most would bother w/.
* In practice, most people pick 1 measure of central tendency + 1 measure of variability
* Notice I’m not getting enough sleep
* We can draw scatterplots to give us a general sense of how closely related 2 variables are.



* Clear the relationship is *qualitatively* the same in both cases: more sleep = less grump!
* Also obvious = relationship between dan.sleep + dan.grump is stronger than baby.sleep + dan.grump
* Plot on the left = “neater” (more condensed around a regression line) than the right.
* To predict my mood, it’d help you to know how many hours my son slept, but it’d be *more* helpful to know how many hours *I* slept.
* Consider these 2 scatterplots:



* Overall strength of the relationships are the same, but the *direction* is different.
* My son sleeps more = I get *more sleep* (positive relationship), but if he sleeps more then I get *less grumpy* (negative relationship)
* Make these ideas a more explicit w/ **Pearson’s correlation coefficient**, traditionally denoted by **r**.
* **Correlation coefficient** between 2 variables X + Y (sometimes denoted **rXY)** = a measure that varies from -1 to 1.
* r = -1 = a perfect negative relationship r = 1 we have a perfect positive relationship
* r = 0, no relationship at all



* The formula for the Pearson’s correlation coefficient can be written in several different ways
* The simplest way to write down the formula is to break it into 2 steps.
* Firstly, introduce the idea of a **covariance** between 2 variables X + Y = a *generalization of the notion of the variance*
* **Covariance =** mathematically simple way of describing the relationship between 2 variables that isn’t terribly informative to humans

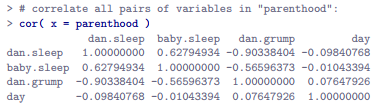


* Just like we saw w/ variance + SD, in practice we divide by N - 1 rather than N.
* B/c we’re multiplying a quantity that depends on X by a quantity that depends on Y + *then* averaging, think of the formula for covariance as an “**average cross product**” between X + Y
* Covariance has the nice property that, if X + Y are entirely *unrelated*, covariance is exactly = 0
* If the relationship between them = positive, covariance is also positive + vice versa
* In other words, the **covariance captures the basic qualitative idea of correlation.**
* Unfortunately, raw magnitude of covariance isn’t easy to interpret + it *depends on the units in which X + Y are expressed*
* Worse yet, actual units of covariance itself are really weird.
* Ex: If X is in units of hours + Y is in units of happiness, the units for their covariance = *hours \* happiness. 🡺 What does that even mean?*
* Pearson correlation coefficient **r** fixes this interpretation problem by ***standardizing the covariance***in the exact same way a z-score standardizes a raw score: *We divide by the standard deviation*.
* However, b/c we have *2* variables that contribute to covariance, the standardization only works if we divide by *both* SDs
* In other words, correlation between X + Y can be written as follows:

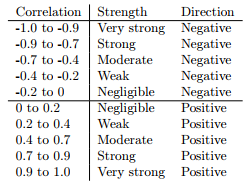


* By doing this standardization, we keep all the nice properties of covariance but the actual values of **r** are on a meaningful scale
* r = 1 implies a perfect positive relationship, + r = - 1 implies a perfect negative relationship.

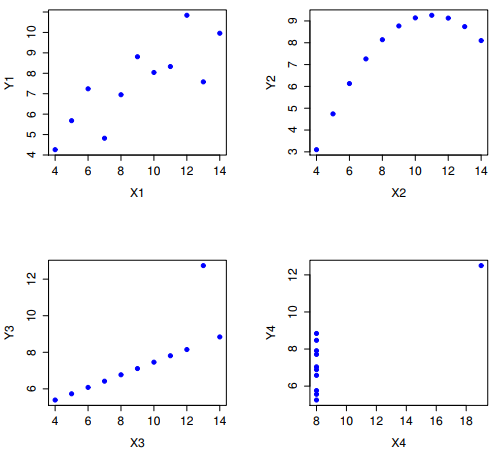
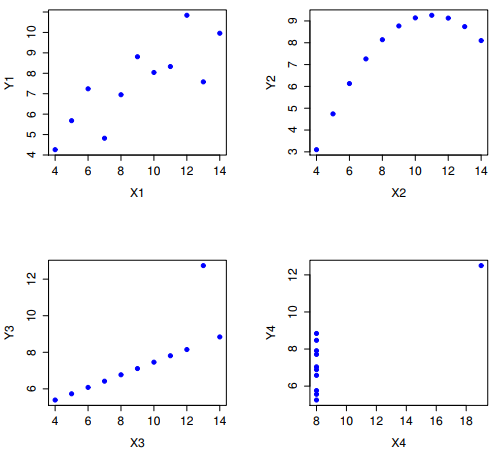




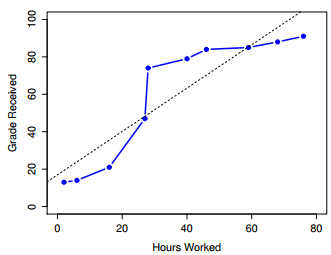
* Naturally, in real life you don’t see many correlations of 1
* How you should interpret a correlation really depends on what you want to use the data for + on how strong the correlations in the field tend to be.
* A friend of mine in engineering once argued that any < .95 is completely useless
* On the other hand there are real cases – even in psychology – where you should really expect correlations that strong.
* For instance, 1 of the benchmark data sets used to test theories of how people judge similarities is so clean that any theory that *can’t* achieve a correlation of *at least .9* isn’t deemed successful.
* However, when looking for, say, elementary correlates of intelligence (e.g., inspection time, response time), if you get a correlation above .3 you’re doing very well.
* ***In short, interpretation of a correlation depends a lot on the context*.**
* Rough Guide:



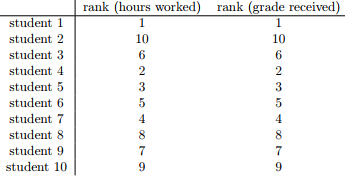
* However, something that can never be stressed enough is that you should **always look at the scatterplot before attaching any interpretation to the data**.
* *A correlation might not mean what you think it means.*
* **Anscombe’s Quartet** = a collection of 4 data sets.
* Each data set 2 variables, X + Y, + for all 4 data sets the mean for X = 9 + the mean for Y = 7.5
* SDs for all X variables are almost identical, as are those for the Y variables, + in each case the correlation between X + Y is r = 0.816.
* You’d think that these 4 data sets would look pretty similar to one another. *They do not.*



* All 4 of these are spectacularly different to each other.
* Lesson here, which so many people seem to forget in real life is **always graph your raw data**
* Pearson correlation coefficient is useful for a lot of things, but it does have shortcomings.
* 1 particular issue stands out: what it *actually* measures = the strength of the *linear* relationship between 2 variables.
* In other words, what it gives you is a *measure of the extent to which the data all tend to fall on a single, perfectly straight line.*
* Often, this is a pretty good approximation to what we mean when we say “relationship” + so the Pearson correlation is a good thing to calculation, and *sometimes, it isn’t.*
* 1 very common situation where the Pearson correlation *isn’t* quite the right thing to use is when an increase in 1 variable X *really is* reflected in an increase in another variable Y, but *the nature of the relationship isn’t necessarily linear*.
* Ex: Relationship between effort + reward when studying for an exam.
* If you put in 0 effort (X) into learning a subject, you should expect a grade of 0% (Y )
* However, a little bit of effort will cause a massive improvement
* Just turning up to lectures means you learn a fair bit, + if you just turn up to classes, + scribble a few things down so your grade might rise to 35%, all w/out a lot of effort.
* However, you just don’t get the same effect at the other end of the scale.
* It takes a lot more effort to get a grade of 90% than it takes to get a grade of 55%.
* What this means is that, if I’ve got data looking at study effort + grades, there’s a pretty good chance that Pearson correlations will be misleading.
* Consider the relationship between hours worked + grade received for 10 students taking some class.



* The curious thing about this is that increasing effort *always increases grade.*
* It might be by a lot or it might be by a little, but increasing effort will *never decrease your grade*.
* Standard Pearson correlation here shows a strong relationship between hours worked + grade received 🡪 r = 0.91
* However, interesting to note there’s actually a perfect **monotonic** relationship between the 2 variables
* In this example at least, increasing hours worked *always increases grade received* (solid line)
* *This* is reflected in a **Spearman correlation** of ρ = 1.
* W/ such a small data set, however, it’s an open question as to which version better describes the actual relationship involved.
* There’s a sense here in which we want to be able to say the correlation is perfect but for a somewhat *different notion of what a “relationship” is.*
* Looking for something that captures the fact that there is a **perfect ordinal relationship** 🡪 *i.e. if student 1 works more hours than student 2, we can guarantee student 1 will get a better grade*
* That’s NOT what a correlation of r = .91 says at all.
* If looking for **ordinal relationships**, *treat the data as if it were ordinal scale*
* So, instead of measuring effort in *terms* of hours worked, rank all 10 students in *order* of hours worked
* That is, student 1 did the least work out of anyone (2 hours) + gets the lowest *rank* (rank = 1).
* Student 4 was next laziest, putting in only 6 hours of work, so they get the next lowest rank = 2



* These data are identical 🡺 The student who put in the most effort got the best grade, the student w/ the least effort got the worst grade, etc.
* We can get R to construct these rankings using **rank()**,

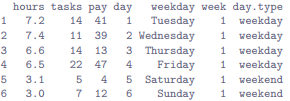
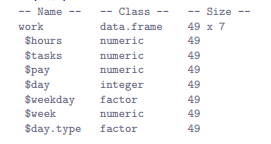




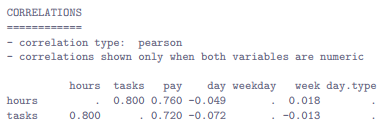
* What we’ve just re-invented is **Spearman’s rank order correlation**, usually denoted **ρ** to distinguish it from the Pearson correlation r.
* We can calculate Spearman’s ρ using R in 2 different ways.
* 1) Do it like above 🡪 rank() to construct rankings + calculate Pearson correlation *on the ranks.*
* 2) Easier to just specify method argument of cor().

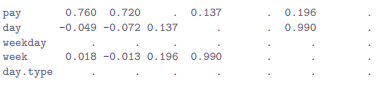


* cor() works pretty well, + handles many of the situations you might be interested in.
* 1 thing many beginners find frustrating: *it’s not built to handle non-numeric variables.*
* From a statistical perspective, this is perfectly sensible as *Pearson + Spearman correlations are only designed to work for numeric variables*
* Suppose you kept track of hours worked in any given day + counted how many tasks you completed.
* If doing the tasks for money, might also want to keep track of how much pay you got for each job.
* It would also be sensible to keep track of the weekday on which you actually did the work
* If you did this for 7 weeks, you might end up w/ a data set that looks like this one:

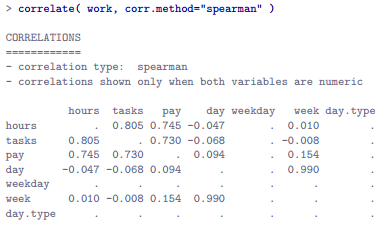


* Obviously, I’d like to know something about how all these variables correlate w/ one another.
* It order to get a correlation matrix, create a new data frame that *doesn’t* contain the factor variables + feed it into cor()
* **correlate()** in **lsr** package can be handy = knows to ignore factors + returns pairwise correlations only between numeric variables:





* See periods whenever a variable is non-numeric + whenever a variable is correlated w/ itself
* **correlate**() can also do Spearman correlations



**Chapter 5.8 - Handling Missing Values**

* Real data sets very frequently turn out to have missing values: perhaps someone forgot to fill in a particular survey question, for instance.
* Missing data can be the source of a lot of tricky issues

**Summary**

* Calculating some basic descriptive statistics is 1 of the very first things you do when analyzing real data, + descriptive statistics are much simpler to understand than inferential statistics
* **Measures of central tendency:** Broadly speaking, central tendency measures tell you where the data are w/ 3 measures typically reported in the literature: the mean, median mode
* **Measures of variability**: tell you about how “spread out” data are w/ key measures: **range, SD, IQR**
* **Standard scores**: z-score = slightly unusual = not quite a descriptive statistic + not quite an inference
* **Correlations**: how strong the relationship is between two variables
* A traditional first course in statistics spends only a small proportion of the class on descriptive statistics, maybe 1 or 2 lectures at most.
* The vast majority of the lecturer’s time is spent on inferential statistics, b/c that’s where all the hard stuff is.
* That makes sense, but it hides the practical everyday importance of choosing good descriptives.
* Good descriptive statistics are descriptive!
* A statistic is an abstraction, a description of events beyond personal experience, + hard to visualize.
* Few if any of us can imagine what the deaths of millions is “really” like, but we can imagine 1 death, + this gives a lone death its feeling of immediate tragedy
* Yet it is not so simple: w/out numbers, counts, or a description of what happened, we have no chance of understanding what really happened, no opportunity event to try to summon the missing feeling.
* Thus it is no small thing to say that the 1ST task of the statistician + the scientist is to summarize data + to find some collection of numbers that can convey to an audience a sense of what has happened.
* This is the job of descriptive statistics, but it’s not a job that can be told solely using the numbers.
* You are a data analyst, not a statistical software package.
* Part of your job is to take these statistics and turn them into a description.
* When you analyze data, it is not sufficient to list off a collection of numbers.
* Always remember that what you’re really trying to do is communicate with a human audience.
* Numbers are important, but they need to be put together into a meaningful story an audience can interpret.
* That means you need to think about framing, context, + about the individual events that your statistics are summarizing