Stats Review

CMPT 353

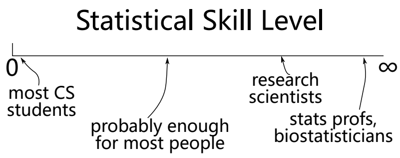
Context

Theme of the course: what do you do to get answers from data? Previous steps: get the data; clean it up until you can work with it.

One possible next step: use statistics to make inferences about what it means.

My goals for this part of the course are fairly modest: you should be able actually *do something* with statistics.

I'll be happy with “probably enough”.



Corollary: if doing statistical analysis wrong would have serious consequences (someone's health, loss of a bunch of money, etc), then ask somebody who knows more statistics than this course will cover.

Let's review some of those things you Definitely Know™ from your prerequisite statistics course…

Types of Data

Quantitative
: Numeric values that have magnitude: -4.2 vs 18.9.

Ordinal
: Ordered values or categories with no magnitude: unsatisfied/neutral/satisfied, 0–9/10–19/20–29/30–39, A+/A/A-/B+/….

Nominal
: Unordered properties or categories: Vancouver/Ottawa, red/green/blue, control/treatment.

We generally think of quantitative data as “data”, but the different categories come up.

Population and Samples

We're usually concerned about the ***population***: all of the values. We want to come to conclusions about the entire population. e.g.

- “Are men taller than women?” ≈ “Is the average height of all men larger than the average height of all women?”
- “Should we put item X on sale?” ≈ “Will we make more profit (from all of our customers' purchasing decisions) if the cost of X was lower than its current value?”

[Figuring out the real question: still not easy.]

But we don't usually get to look at the entire population (especially if it's extremely large or infinite). We usually have to deal with just a ***sample***: a subset of the population. e.g.

- 50 men and 50 women: measure their heights.
- A fraction of customers who are offered the lower price, compared to the rest.

The point of inferential statistics is to use (well-chosen) samples to come to (probably-correct) conclusions about the population.

- Yes, the average height of men is larger than the average height of women.
- No, don't put X on sale: we think it will make less money.

Probability Distributions

If we have some random thing happening (a ***random variable***, like sampling an individual from a population), what is the probability of a certain outcome? (e.g. height = 1.80 m, heads/tails).

A ***probability distribution*** is the description of probabilities for all outcomes.

A ***discrete probability distribution*** has outcomes from a discrete (usually finite) set. e.g. flipping a coin, number of times a Wikipedia page will be viewed tomorrow.

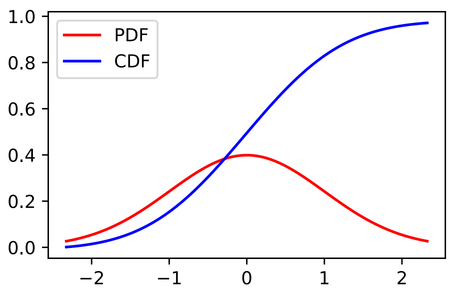
Of course, the sum of probability of every possible outcome must be one.

$$\sum_{u \in U} P(u) = 1$$

For a discrete or continuous probability distribution, we can talk about the ***cumulative distribution function***, the probability of the outcome being less-than-or-equal-to a particular value. Often written $F(x)$.

Or its derivative, the ***probability density function***, often $f(x)$.

The picture is probably more useful. A normal distribution's cumulative distribution function, and probability density function:



Central Tendency

Where is “the center” of your data?

Most commonly used: the ***mean*** or ***expected value***.

For the population, these are the same, and usually called μ or $E(X)$.

A sample has a mean (but no expected value), $\overline{x}$.

The sample mean $\overline{x}$ is an ***unbiased estimator*** of the population mean μ: if the sample is randomly chosen, $E(\overline{x}) = \mu$.

i.e. If you take a good sample and calculate the mean, you have a meaningful estimate the population mean.

Can also look at the ***median***: value in the middle when sorted; 50th percentile.

Or the ***mode***: value that occurs most frequently.

Dispersion

How spread out is the data? How far away from the mean are the points "usually" found?

Most commonly used: ***standard deviation*** of a population σ, or a sample s. Or ***variance***: σ^2 or s^2.

Again, the sample standard deviation is an unbiased estimator of the population standard deviation: $E(s) = \sigma$.

In general, at least half of a population is within $\sigma\sqrt{2}$ of the mean. For a normal distribution, ≈68% is within σ.

So *mean* is something like "where is the middle of the data?" The *standard deviation* is "how spread out is the data from the mean?"

Pandas can make quick work of showing you summary stats for a DataFrame:

```
print(data.describe())
```

```
                 id       rating     timestamp
count  1.669000e+03  1669.000000  1.669000e+03
mean   8.245327e+17    11.762133  1.485419e+09
std    9.829935e+16     1.646146  2.343639e+07
min    6.989080e+17     0.000000  1.455468e+09
25%    7.486928e+17    11.000000  1.467337e+09
50%    8.026004e+17    12.000000  1.480190e+09
75%    8.834828e+17    13.000000  1.499474e+09
max    1.125920e+18    17.000000  1.557275e+09
```

Pandas doesn't know that "id" is actually nominal, so that column isn't meaningful.

Relationships

With two or more variables, how are they related?

The ***covariance*** ($cov(X, Y)$ or $\sigma_{X,Y}$ for populations, $s_{X,Y}$ for samples) gives information about the joint variability: do they change together or independently?

Note: s_X is sample standard deviation and s_X^2 is variance, but $s_{X,Y}$ is sample covariance

Positive covariance: larger Y usually happen with larger X. Negative covariance: larger Y usually happen with smaller X.

The ***correlation coefficient*** is basically the same info, but normalized into a -1 to 1 range. ρ (rho) for populations, r for samples.

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$
$$r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y}$$

Values close to -1 and 1: a lot of **linear** relationship between the variables. Close to 0: little or no linear relationship.

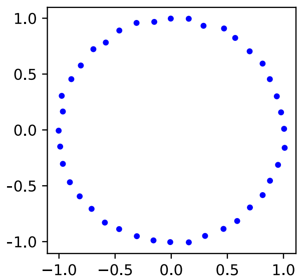
If you're interested in two variables' correlation coefficient, we also have a quick tool for that:

```
print(stats.linregress(data['timestamp'], data['rating']).rvalue)
```

```
0.5005674118565123
```

Remember that $r \approx 0$ for some data means there's no apparent *linear* relation, not that x and y aren't related to each other.

This data has $r = 0.003$, but has a fairly obvious relationship between x and y.

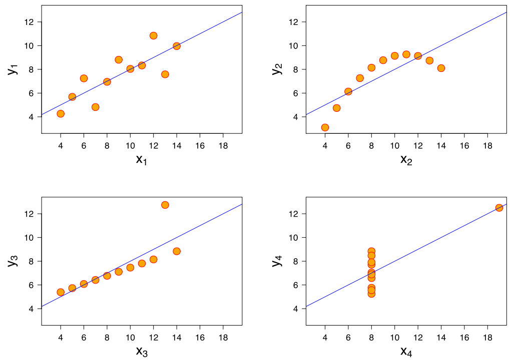


Plotting Data

The basic summary stats can only tell you so much: they are often not enough information to really understand what's happening in a data set.

You should start by plotting your data, even if you don't “need” a plot. It can tell much more of a story.

Anscombe's quartet: four data sets each with $\overline{x} = 9.00$, $\overline{y} = 7.50$, $s_x^2 = 11.00$, $s_y^2 = 4.125$, $r_{xy} = 0.816$, regression line $y = 3.00 + 0.500x$. [*]

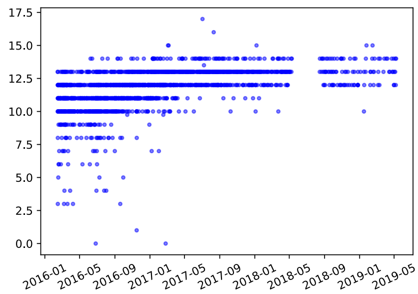


The lesson: start by making a quick plot of your data.

It can tell you a lot that basic summary stats can't. (But summary stats can often tell you things a plot can't.)

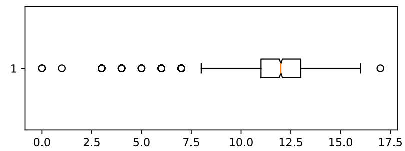
[Also, outliers can really affect your analysis.]

For two-dimensional data, a scatter plot is often the most obvious: shows you the approximate distribution of both variables, and any obvious relationships.



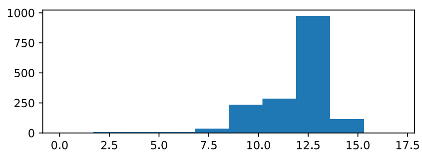
For one variable, maybe a box plot. Shows the median, quartiles, range, outliers.

```
plt.boxplot(data['rating'], notch=True, vert=False)
```



Or a histogram: a look at the rough shape of the distribution.

```
plt.hist(data['rating'])
```

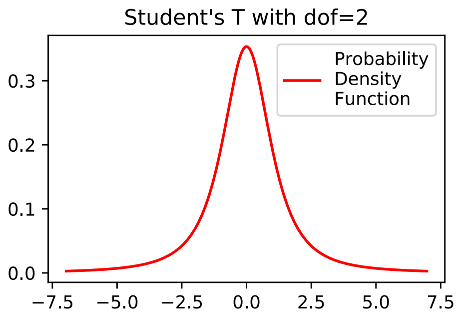


Specific Distributions

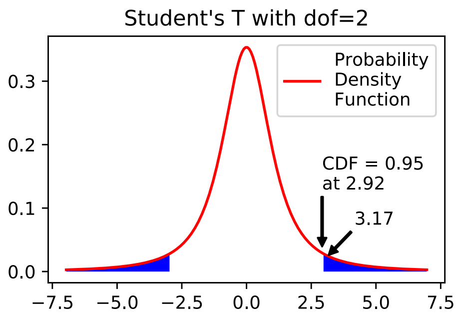
There are several probability distributions that are often interesting.

Some of them are used to come to a conclusion while inferring something about the data: more later.

For example, we may come up with a random process and realize "if H is true and we do [a bunch of arithmetic], then we get a value sampled from a 'Student's T' distribution."



What if we do that and find a value of 3.17? There's a <10% probability of sampling a value that far from the mean, so it seems unlikely that H is true.



... maybe that's useful to notice.

Normal Distribution

One in particular that comes up a lot: the ***normal distribution***, generally written $\mathcal{N}(\mu, \sigma^2)$.

e.g. for Kalman filters, we assumed the noise was normally-distributed (with $\mu = 0$ and we guessed σ^2 as closely as we could).

It's very common to get normally-distributed values when doing random sampling.

e.g. flip n coins: number of heads is distributed $\mathcal{N}(\frac{n}{2}, \frac{n}{4})$ (if n is large).

The central limit theorem (more later) says that you can get a normal distribution anywhere if n is large enough, and you look at your data the right way.