Stats Review

EC420 MSU

Justin Kirkpatrick
Last updated January 10, 2020

This lecture

MICHIGAN STATE UNIVERSITY

Goal: Refresh or catch up on statistical concepts relevant to econometrics.

1. Selected questions from D2L
2. Random variables
3. Sample vs. Population
4. Probability density functions (PDF/PMF, CDF)
5. Mean, variance (in population, in sample)
6. Correlation, independence, mean independence, iid
7. Types of distributions (normal, uniform, student-t, χ^2, bernoulli, poisson)
8. Statistical inference
9. Useful functions and properties

Questions from last week

Class Business

MICHIGAN STATE UNIVERSITY

Some of the most common questions re: the class from your cards on Monday

- Will we go over R?
  - Most definitely. I will post instructions on installing R and Rstudio, and you will take a tutorial as part of your first problem set.
- How difficult will problem sets be?
  - Moderately difficult. If you're familiar with the concept of computer scripting, some parts will be easier. If you're not, that's **100 percent OK**, you do **not** need to have any R experience for this class. One of the points of the class is to get that experience.
- Will R appear on the quiz/exam?
  - In very limited form - at most, questions about which command to use to solve a stated problem.

- How often will we use t-tables?
  - Once.
- How detailed do the reading responses need to be?
  - Just enough to show some thought on the topic. Part of the purpose of the RR is to give you a chance to think through the concepts, so you just have to show some original thought on the matter, even if that thought is confusion. The other part is to ensure you've done the reading, so at least show me some attempt to digest the reading.
- I have class during Office Hours
  - We have a TA who is quite good and will be very helpful (likely Thursdays 1-2:45pm), but you can also email me to set up hours outside of office hours.

- Are there parts of the class helpful for experimental research?
  - The concept (and purpose) of randomization is important in experimental research, so that is probably the largest overlap. Also, data management and manipulation is a big part of all quantitative research. We will also do a little bit of visualization.
- Will it be mostly coding?
  - No. The statistical concepts are the priority, R and coding is secondary but still useful and necessary.
- What was the average grade last semester?
  - Around a 2.5, I think.

I couldn't get to all the questions, so if I didn't cover yours but want to talk about it further, come to Office Hours or setup a meeting via e-mail.

Stats Review!

Random Variables

Wooldridge defines a random variable as

"*...one that takes on numerical values and has an outcome that is determined by an experiment.*"

When referring to a random variable (RV), we use an upper-case e.g. X

RV's have *realizations* ("...determined by an experiment")

- Like flipping a coin
- Or rolling a die

We label these using lower-case: x

And when we have multiple realizations of a RV, we can label them:
$\{x_1, x_2, x_3\} = \{6, 1, 5\}$

And when we have a map of how this RV behaves, we have a **distribution** e.g.
$X \sim N(0, 1)$

- E.g. the "normal" distribution.
- One important element of a distribution is the *support*, which is the possible values that the RV can take

As an example, maybe X is a roll of a normal die

- The support of X is $\{1, 2, 3, 4, 5, 6\}$
- The distribution of X is:
  - $Pr(X = 1) = \frac{1}{6}$
  - $Pr(X = 2) = \frac{1}{6}$
  - $Pr(X = 3) = \frac{1}{6} \cdots$
- The random variable is not the realizations
  - "The map is not the territory"

Discrete vs. Continuous

A random variable is discrete if it can take on only a finite or countably infinite number of values.

- Bernoulli (coin flip) can take on only two values, $\{0, 1\}$. It is discrete.
- Poisson (count) can take on the values $\{0, 1, 2, \dots\}$, which is countably infinite. It is discrete.
- Normal can take on any value $\in \{-\infty, +\infty\}$*
- Whether or not a RV is discrete or continuous is determined by it's support

* $\in$ = "in"

Population

The "population" is what we'd like to learn about.

Let X be a random variable (RV) representing a population

- X could be "hourly wage"
- X could be "age"

There is some feature of the population we are interested in

- The mean hourly wage
- The distribution of age
- This is the "population statistic"

But we cannot collect complete data on the population:

- It may be too expensive or it just isn't possible

So we have to take a *sample*

For example:

When we do national Presidential polling, we are *trying* to measure the national vote. The RV here is "voting preference"

- Call the true fraction of all people who intend on voting for Candidate A the "population statistic"
- But we do not get to call every person in America
- Even if we had infinite phones and infinite interns, the *population* we're interested in might not be that easy to figure out
  - "Likely voters"
- So we instead settle for getting as close to the population statistic as we can. We use a sample.

Population Effect

- In econometrics, we're often going to be concerned about a population *effect*:
  - Effect of a drug on a health outcome
  - Effect of a work training program
- Just as there is a population statistic we can call "average vote", there is a "population effect" of a drug or work training program.
  - Just as we can't call everyone in the country about their vote, we can't test everyone in the country for a drug's effect
- We don't observe the whole population, so we cannot directly calculate the population statistic we're interested in.

Sample

The sample is what we observe.

It is always taken from the population. We use N (or n) to denote it's size.

And under certain conditions* the sample tells us about the population

Population statistics will **always get a greek letter**. Sample statistics don't.

- μ is the population mean
- σ^2 is the population variance
- $\bar{x}$ is the sample mean
- s^2 is the sample variance

* We will learn a lot about these conditions this semester.

Sample vs. Population MICHIGAN STATE UNIVERSITY

For any *population statistic* we are interested in, there will be a *sample statistic* that will tell us something about the *population statistic*.

- In the voting example, our sample poll tells us *something* about the population

The sample statistic will never be exactly the population statistic

But it will be close

The statistical applications of econometrics are all about quantifying "how close"

In everything we do for the first half, we will use the concept of "population" and "sample" heavily.

Please make sure you are comfortable with the difference.

Any questions?

Probability density functions

Probability density function

The *probability density function* (pdf) of X summarizes the information concerning the possible outcomes of X and the corresponding probabilities in the population

$$f_X(x) = \begin{cases} p_j \text{ if } x = x_j \\ 0 \text{ otherwise} \end{cases}$$

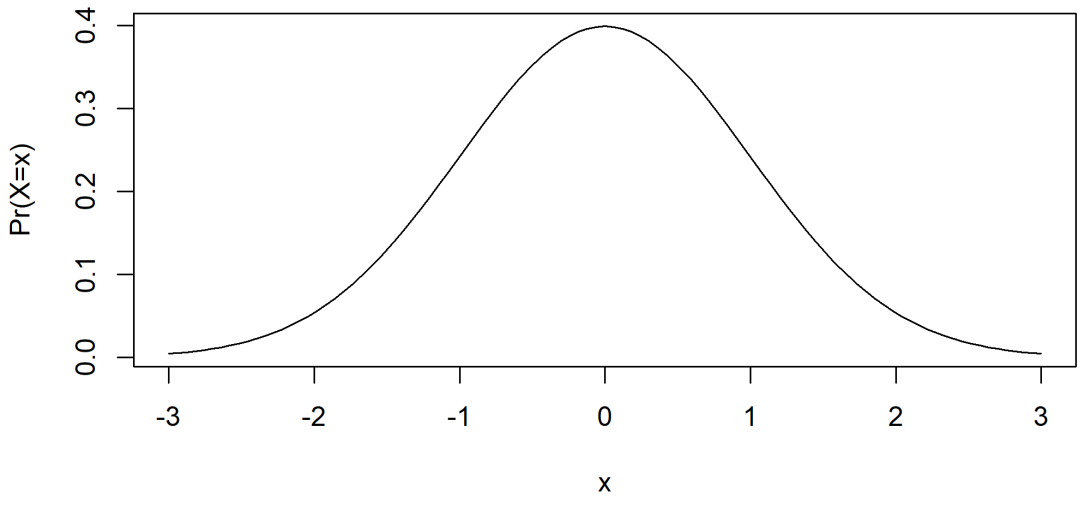
The *pdf* tells us the probability of realizing any value within a very small window around the value. If we know the distribution, including the parameters, then we can plug in any value.

In a Normal $X \sim N(\mu, \sigma^2)$, then $f_X(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$

- Just note that the pdf pertains to a RV, X, and some number you want to plug in, here labeled t.

A *pdf* must **always sum (discrete) or integrate (continuous) to one**

A PDF



The *cumulative density function*, or *cdf*:

The *cdf* tells us the probability of a RV realization being less than some value:

$$Pr(X \leq t) = F_X(t) = \int_{-\infty}^{t} f_X(s)ds$$

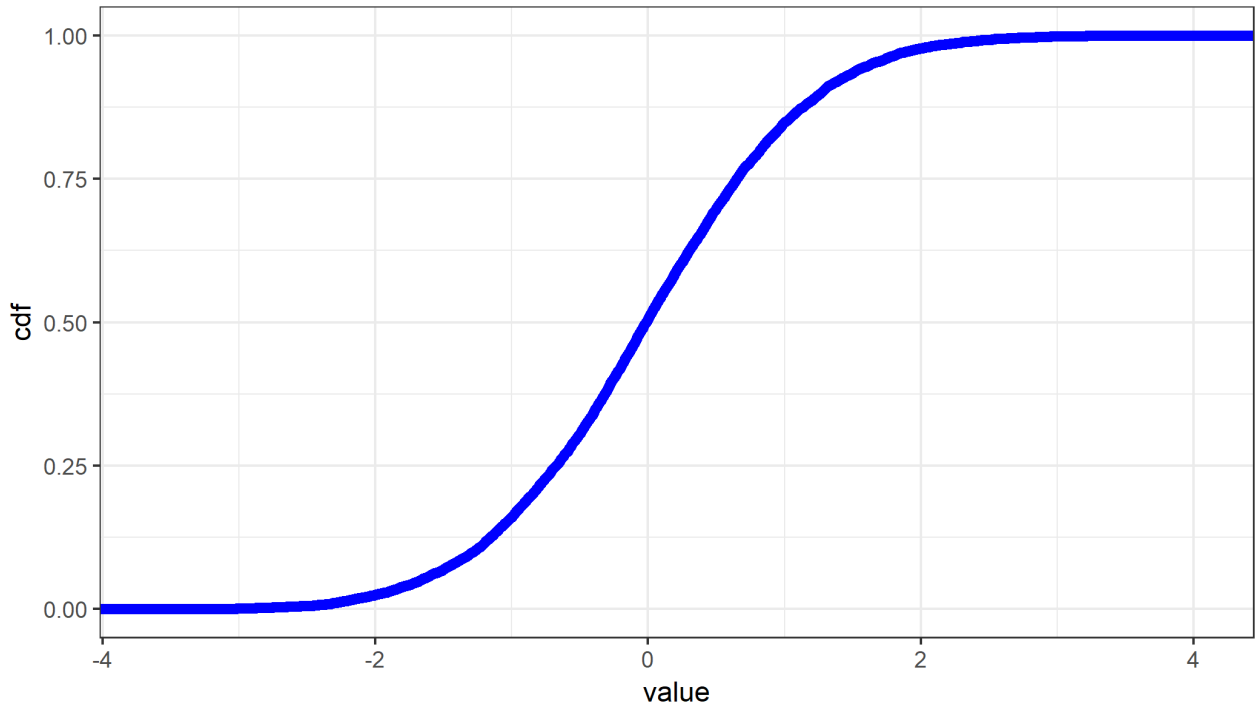
The *cdf* has a few useful properties:

The probability of a RV X being **greater than** some value t is $1 - F_X(t)$

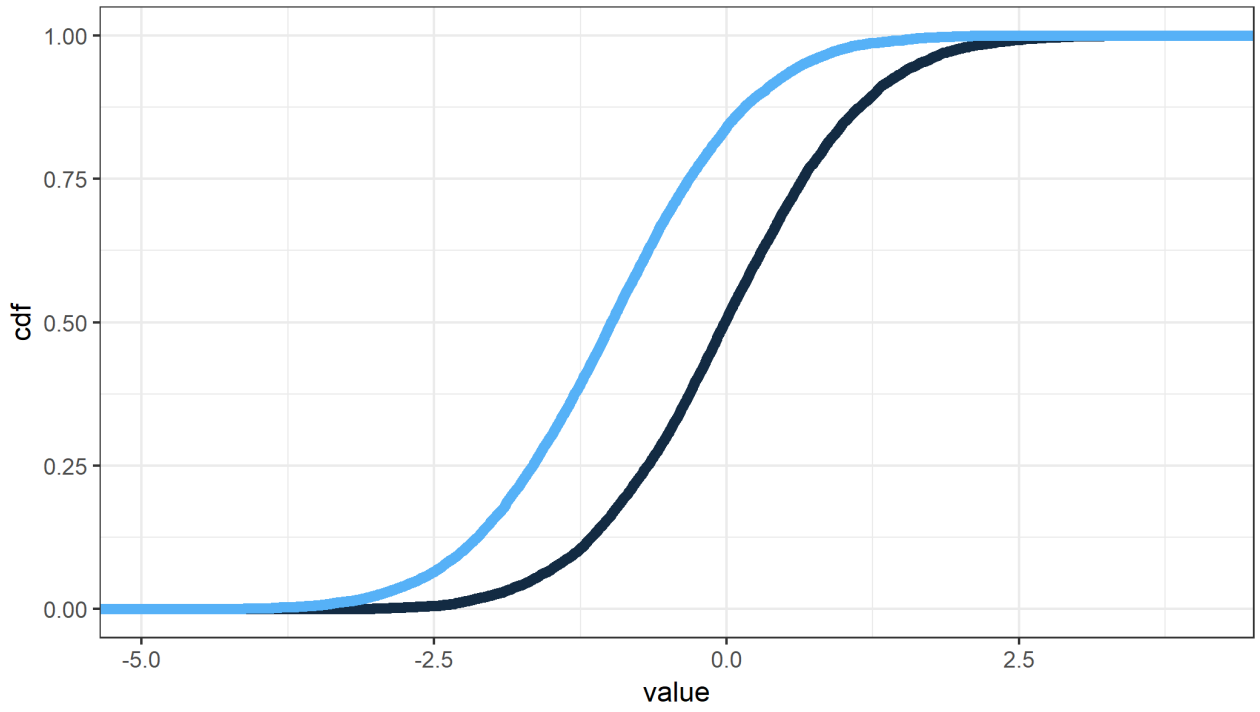
The derivative of the *cdf* at any value t is the *pdf* at t

The *cdf* is always $\in [0, 1]$

The *cdf* is denoted as F, while the PDF is f.



Another example



PDF and CDF

The PDF and the CDF describe the RV. Therefore, they are statements about the population.

- They are also not directly observed
- We may say that $X \sim N(0, 1)$
  - "X is distributed Normal with a mean of 0 and a variance of 1"
  - But we never get to know this for sure in practice

A probability distribution (and thus an RV described by a distribution), has "moments"

- Mean
- Variance
- And more!

Let's look at those. But first, we'll define their sample equivalents.

Mean, variance

Sample mean

It should surprise nobody in this room that the *sample mean* is calculated as:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{N} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

Note that N is the sample size, n is the index, and x_n are the realizations of X

$\bar{x}$ is a *sample statistic* and a *descriptive statistic* (it describes the sample $x_1, \ldots, x_n$). We will also see that it is an *estimate* of the population mean, μ_X. Let's hold off on this for now.

Property of the sample mean (and summation operator)

Remember that multiplying the sample by a constant means you can take the constant out:

$$\frac{1}{N}\sum_{n=1}^{N} cx_n = c\bar{x}$$

But you cannot do this with the product of two RVs

$$\frac{1}{N}\sum_{n=1}^{N} x_n y_n \neq \bar{x}\bar{y}$$

You can move the addition or subtraction of a constant in or out:

$$\frac{1}{N}\sum_{n=1}^{N}(x_n - c) = \bar{x} - \frac{1}{N}Nc = \bar{x} - c$$

Just as we cannot say that the mean of the product of two RV's is the product of the means, we cannot say that the mean of one RV squared is the mean squared

$$\frac{1}{N}\sum_{n=1}^{N} x_n^2 \neq \bar{x}^2$$

That is, the average of the square of a sample is **not** the square of the average

If we had a sample from X of $\{1, 2, 6\}$, then $\bar{x} = \frac{1+2+6}{3} = 3$

But $\frac{1+4+36}{3} = 13.6667$ which is **not** 9.

The *sample average*, $\bar{x}$, can be treated like a constant since it is the realization from a sample.

This is important as we will be working with deviations from the mean a lot:

$$\frac{1}{N}\sum_{n=1}^{N} x_n - \bar{x}$$

ClassID	Score
a	4
b	5
c	3
d	5
e	5
f	2
Avg	4

One particularly important property is revealed here: the sum of all deviations from the *sample mean* is zero.

$$(4-4)+(5-4)+(3-4)+$$
$$(5-4)+(5-4)+(2-4)=0$$

$$\frac{1}{N}\sum_{n=1}^{N} x_n - \bar{x} = \bar{x} - \bar{x} = 0$$

Expected value

The expected value is a very similar concept to the mean. When refering to an RV, "mean" and "expected value" are interchangable.

Expected value is a population concept, while the sample mean exists only for any set of *realizations* of a RV.

The *expected value* of a RV is the "measure of central tendency"

- A fancy way of saying "tells us about the middle"
- Sometimes called the "first moment"

It is represented by the E operator and is read as "the expectation of X". In a discrete RV, the Expectation is defined by:

$$E(X) = x_1 f(x_1) + x_2 f(x_2) + \ldots + x_n f(x_n)$$

Remember, f is the pdf, and is a population concept

For continuous RVs:

$$E(X) = \int_{-\infty}^{+\infty} x_s f(x_s) ds$$

This is the integral over all possible values of x, weighted by the pdf at that value.

Note that we don't use a sample to calculate this - we need to know the support (the values X can take) and the pdf, $f(x)$.

Expectation operator

$E(X)$ has some useful properties (similar to the mean), where c is a constant (not a RV):

- $E(cX) = cE(X)$
- $E(cX + b) = b + cE(X)$

That is, you can add and multiply constants in and out of the expectation

Expectation operator

But in general:

- $E(XY) \neq E(X)E(Y)$ in general
  - This **will** be true in certain cases, which will be **very** useful to us!
- $E(X^2) \neq [E(X)]^2$

Conditional Expectation

In an upcoming lecture, we will think about conditional expectations:

- $E(X|Y = y)$ which is the expectation of X *conditional on* Y taking the value y

We will work a lot more with the expected value in this class

Mean, variance MICHIGAN STATE UNIVERSITY

Expected value continued...

The *expected value* will also be the best guess of a realization of a RV

- "best" as in "minimizes the sum of differences between realizations and the guess"
- If I have to pay $1 for every unit off I am from a realization of a RV, I am best-off by guessing the expected value
- This applies for any monotonically increasing loss function
  - Loss function is any way of calculating a penalty
  - For example, it could be the square of the difference
  - Or, the absolute value of the difference
  - "Increasing" means its result increases with the input
  - This excludes *The Price Is Right* pricing

Expected value and (population mean) are essentially interchangable

$$E(X) = \mu_X$$

Both are population concepts, so these are not observed

- Thus, the greek letter for μ
- The extra X in μ_X just clarifies what RV's population mean we're referring to

Variance

variance is a "measure of diffusion"

- A fancy way of saying "how spread out are the likely values"
- Sometimes called the "second moment"
- Not all RVs have a finite variance; some are ∞
  - All samples do - you can always calculate the variance from a finite number of draws
  - ...but not all RVs have a finite variance

Most important, two RVs or two samples of two RVs can have the same mean but different variance

- The mean is still the "best guess"
- The variance tells us how far off your "best guess" will be on average
- Higher variance = more spread = further off

Variance, in a sample, is calculated by summing the squared differences between the realizations of X and the mean of X. The concept of variance can refer to the population (population variance) **or** the sample (sample variance)

It is written as $Var(X)$ or σ_X^2 (or s^2 for sample variance)

If we know the whole population and the population mean, we would use:

$$Var(X) = \sigma_X^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2$$

But we rarely know μ or observe the whole population.

But we do have a good sample analog using $\bar{x}$!

Sample variance

So, we can get a **sample variance**, s^2, by first calculating $\bar{x}$, taking it as a constant, and calculating:

$$s^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_i - \bar{x})^2$$

Note the $N-1$ in the denominator! This corrects for the fact that $\bar{x}$ is an estimate

We can write the variance using expectations as well:

(Note that I'm using the population variance here)

$$\begin{aligned}
\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2 &= \frac{1}{N}\sum_{n=1}^{N}\left[x_n^2-2x_n\mu+\mu^2\right] \text{ (Expand polynomial)} \\
&= \frac{1}{N}\sum_{n=1}^{N}x_n^2-\frac{1}{N}\sum_{n=1}^{N}2\mu x_n+\frac{1}{N}\sum_{n=1}^{N}\mu^2 \text{ (Distribute the sum} \\
&= \frac{1}{N}\sum_{n=1}^{N}x_n^2-2\mu\frac{1}{N}\sum_{n=1}^{N}x_n+\frac{1}{N}N\mu^2 \text{ (Move constants out)} \\
&= E(X^2)-2\mu E(X)+\mu^2 \\
&= E(X^2)-2E(X)E(X)+[E(X)]^2 \\
&= E(X^2)-[E(X)]^2
\end{aligned}$$

Operations on Variance where a and b are constants differ from the Expected Value operations:

$$Var(aX+b)=a^2Var(X)+0$$

- Constants have zero variance.
- Scaling X by a scales variance by a^2.
- This **will** be important in this class.

If we have an RV, and we know it's population mean and variance: μ and σ^2

We can standardize X by subtracting the mean and dividing by the standard deviation:

$$Z = \frac{X - \mu}{\sigma}$$

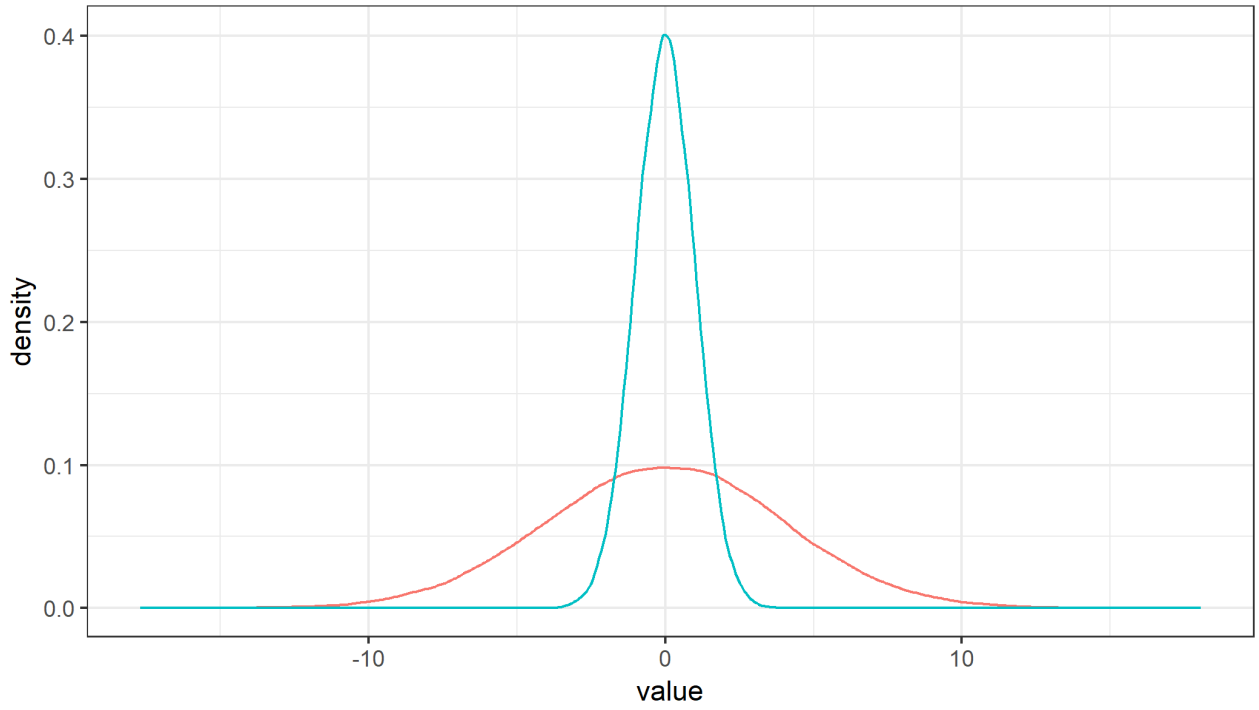
Write a as $\frac{1}{\sigma}$ and b as $-\frac{\mu}{\sigma}$. Then

$$Z = \frac{1}{\sigma}X + -\frac{\mu}{\sigma} = aX + b$$

$$E(Z) = aE(X) + b = \frac{\mu}{\sigma} - (\frac{\mu}{\sigma}) = 0$$

$$Var(Z) = a^2 Var(X) + 0 = \frac{1}{\sigma^2}\sigma^2 = 1$$

Low variance is $\sigma^2 = 1$; high variance is $\sigma^2 = 16$. Both means are zero.



Covariance, correlation

We may be interested in how **two random variables** behave together.

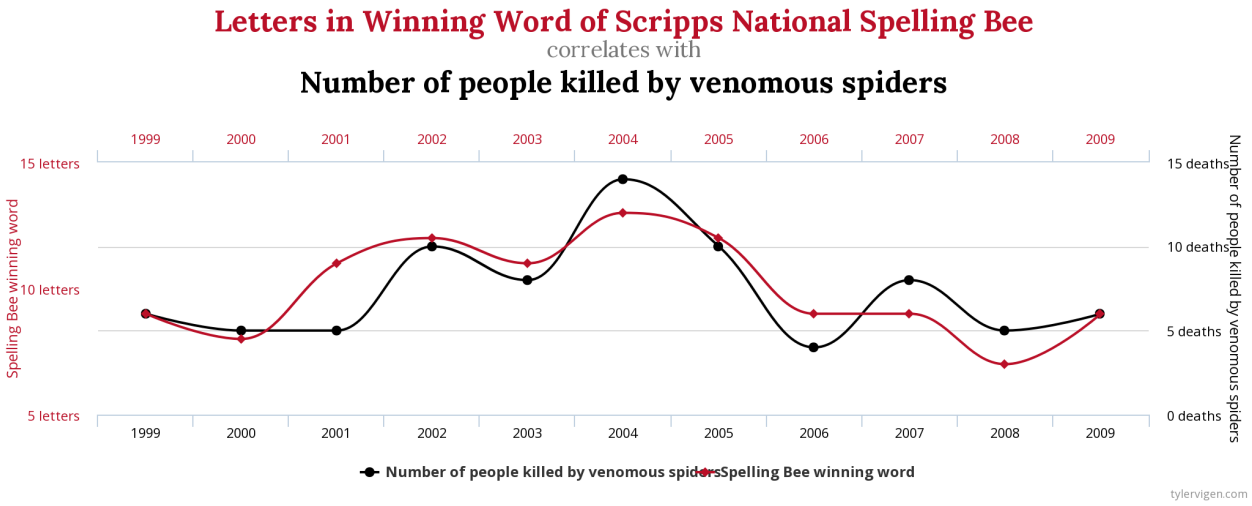
- Income and age
- Wage and education
- Corn yields and fertilizer
- Snowfall and traffic fatalities
- "Letters in Winning Word in Scripps National Spelling Bee" and "Deaths from Venemous Spiders"

Each of these are *pairs* of RV's.

Just as we have measures of central tendency and dispersion, we will have a measure of this association betwen RV's: **covariance**.

And we can also express the relationship between the PDF's of the RV's.

Yes, it's real



Let's define **Covariance** between X and Y

$$Cov(X,Y) = \sum_{n=1}^{N}(x_n - \mu_X)(y_n - \mu_Y)$$

Note that we are *summing the pairwise deviations from the mean.*

Covariance will be:

-**higher** if x is above the mean when y is also above the mean

-**lower** if x is below the mean when y is also below the mean

-**zero** if x is randomly above/below the mean when y is above/below the mean

Covariance is a measure of how closely two RV's track each other.

With some algebra, we can also write covariance as:

$$Cov(X,Y) = \sum_{n=1}^{N}(x_n - \mu_X)(y_n - \mu_Y) = E(XY) - \mu_X\mu_Y$$

Just as $E(X)$ was important to the *measure of central tendency* and $E(X^2)$ was important to the *measure of dispersion* (variance), $E(XY)$ is important to the covariance, a *measure of association*.

We can also scale the covariance so that it is between -1 and 1:

$$\frac{Cov(X,Y)}{Var(X)Var(Y)} = \rho \in [-1,1]$$

We will use ρ for the *correlation coefficient*, though you'll frequently see ρ used for other purposes as well.

With regards to the elements of $\rho : Cov(X,Y), Var(X), Var(Y)$, in what case would:

- $\rho = 1$?
- $\rho = -1$?
- $\rho = 0$

This is a good stopping point