Statistical Inference Notes - WEEK 1

Statistical Inference is the process of drawing formal conclusions; as settings where one wants to infer facts about a population using noisy statistical data where uncertainty must be accounted for.

A strong inference may affect the study itself. Researchers may decide it is ethically responsible to halt the study in light of strongly positive or negative preliminary results.

Considerations of a study include:

- Is the sample representative of the target population?
- Is the sample representative of the target population?
- Are there known and observed, known and unobserved, or unknown and unobserved variables that contaminate our conclusions?
- Is there systematic bias created by missing data or the design or conduct of the study?
- What randomness exists in the data and how do we use or adjust for it?

  Randomness can be explicit via randomization or random sampling, or implicit as the aggregation of many complex unknown processes.
- Are we trying to estimate an underlying mechanistic model of phenomena under study?

Example goals of Inference:

- Estimate or quantify the uncertainty of an estimate.
- Determine whether the quantity is a benchmark value (”Is the treatment effective?”)
- Infer a mechanistic relationship when quantities are measured with noise (”What is the slope for Hooke’s Law?”)
- Determine the overall impact of a policy a phenomenon, as opposed to revealing the mechanism that describes it.

Some tools of the trade:

- Randomization: balances unobserved variables that may confound inferences, i.e between the control and treated.
- Random Sampling: data obtained is representative of the population of interest.
- Sampling models: a model for the sampling process is created because Randomization and Random Sampling are often impossible. ”iid”.
- Hypothesis Testing: Decision making in the presence of uncertainty.
- Confidence Intervals: Quantify the uncertainty in estimation.
- Probability Models: A formal connection between the data and the population of interest; assumed or approximated.
- Study Design: Designing the experiment to minimize bias and variability. Of course randomized is the best.
- Nonparametric Bootstrapping: The process of using the data (with minimal Probability Model assumptions) to create inferences.
- Permutation, Randomization, and Exchangability Testing: The process of using data permutations to perform inferences.

Thinking Styles:

Data scientists use some combination of two opposing modes of inference (as well as other schools of thought):

- Frequency Inference: What should I decide given the long-run-proportion of events in independent, identically distributed repetitions?
- Bayesian Inference: Given my subjective beliefs, and the new objective information from the data, how should I modify my beliefs?

This class will focus on frequency style analyses, beginning with probability modeling.

Probability

Notation

- Ω: The sample space; the collection of all possible outcomes.

  e.g. die roll: $\Omega = \{1, 2, 3, 4, 5, 6\}$

- E: (Or another letter.) An event; a subset of Ω.

  e.g. die roll is even: $E = \{2, 4, 6\}$

- ω: (Or another letter.) An elementary or simple event is a particular result of an experiment.

  e.g. die roll is a four: $\omega = 4$

- $\emptyset$: The null event or empty set.

Common Set Operations

- $\omega \in E$: The simple event ω is an element of set E.

  Implies that E occurs when ω occurs.

  e.g. If I rolled a 4, then I rolled an even number.

- $\omega \notin E$: The simple event is not an element of the set.

  Implies that E doesn't occur when ω occurs.

- $E \subset F$: The set E is a subset of set F.

  The occurrence of E implies the occurrence of F.

- $E \cap F$: Intersection; the set containing the elements that are common to both E and F.

  Implies an event for which both E and F occur.

- $E \cup F$: Union; the set containing all the elements in both E and F.

  Implies an event for which E or F or both occur.

- $E \cap F = \emptyset$: E and F are mutually exclusive; both cannot occur; no intersection.

- E^c or $\bar{E}$: The event for which E doesn't occur.

Probability

A probability measure, P, is a function from the sample space Ω for which the following hold true:

1. For an event $E \subset \Omega, 0 \leq P(E) \leq 1$.

   For an event in a subset of the sample space,

   the probability is between 0 and 1.

2. $P(\Omega) = 1$.

   The probability of an event in the entire sample space is 1.

3. If $E_1 \cap E_2 = \emptyset, P(E_1 \cup E_2) = P(E_1) + P(E_2)$.

   If E_1 and E_2 have no intersection,

   then the probability of their union is the sum of their probabilities.

4. #3 implies finite additivity:

   $$P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$$

   where the A_i are mutually exclusive.

   If you union up a bunch of mutually exclusive events,

   then the union can become a sum.

Example Consequences

Andrey Nikolaevich Kolmogorov in 1933 formulated these eight axioms and said that these are all you need to have probability behave as we think it should.

- $P(\emptyset) = 0$

  also $P(\Omega) = 1$

- $P(E) = 1 - P(\bar{E})$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- if $A \subset B$ then $P(A) \leq P(B)$
- $P(A \cup B) = 1 - P(\bar{A} \cap \bar{B})$

- $P(A \cap \bar{B}) = P(A) - P(A \cap B)$
- $P(\cup_{i=1}^{n} E_i) \leq \sum_{i=1}^{n} P(E_i)$
- $P(\cup_{i=1}^{n} E_i) \geq \max_i P(E_i)$

Random Variables

1. discrete: $P(X = k)$
2. continuous: $P(X \in A)$

PMF - Probability Mass Function

The probability mass function describes a discrete random variable. A PMF evaluated at a value corresponds to the probability that a random variable takes that value. To be a valid PMF, a function p must satisfy:

1. $p(x) \geq 0$ for all x.
2. $\sum_x p(x) = 1$

where the sum is taken over all possible values for x.

Example 1:

Let X be the result of a coin flip where $X = \{0, 1\}$ or {heads,tails}.

$$\begin{aligned} p(x) &= (1/2)^x (1/2)^{1-x} \text{ for } x = 0, 1 \\ p(1) &= (1/2)^1 (1/2)^{1-1} \; = \; 1/2 \\ p(0) &= (1/2)^0 (1/2)^{1-0} \; = \; 1/2 \end{aligned}$$

Example 2:

Suppose the coin isn't fair. Let θ be the probability of a head expressed as a proportion (between 0 and 1), say 0.25.

$$\begin{aligned} p(x) &= \theta^x (1-\theta)^{1-x} \text{ for } x = 0, 1 \\ p(1) &= \theta^1 (1-\theta)^{1-1} \; = \; \theta \\ p(0) &= \theta^0 (1-\theta)^{1-0} \; = \; 1-\theta \end{aligned}$$

So if heads is $\theta = 0.25$, then tails is $1 - 0.25 = 0.75$.

PDF - Probability Density Function

The probability density function describes a continuous random variable. The area under the PDF line corresponds to a probability for a group of values of the variable.
To be a valid PDF, a function f must satisfy:

1. $f(x) \geq$ for all x.
2. The area under $f(x)$ is one.

A single value of a continuous random variable corresponds to an area of 0. Instead, we refer to a probability of a region of values (between a and b). The fact that the probability of a single specific value is 0 is a consequence of modeling the probability as a truly continuous entity, an infinite decimal expansion, wherein we cannot attain infinite precision. In other words, there is no such thing as a single specific value.

The Gaussian Bell Curve is the most famous example of a probability density function.

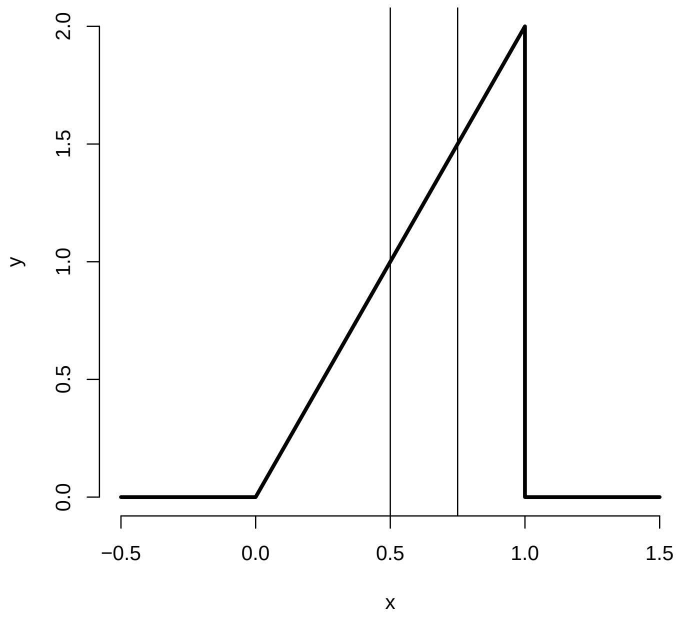
Example 1:

Suppose that the proportion of help calls that get addressed in a random day by a help line is given by this probability density:

$$f(x) = \begin{cases} 2x & \text{for } 1 > x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Is this a mathematically valid density?

```
> x <- c(-0.5, 0, 1, 1, 1.5)
> y <- c(0, 0, 2, 0, 0)
> #par(pin = c(2,2), mar = (c(2, 2, 2, 1) + 0.1))
> plot(x,y, lwd=3, frame=F, type="l")
> abline(v=0.5)
> abline(v=0.75)
```



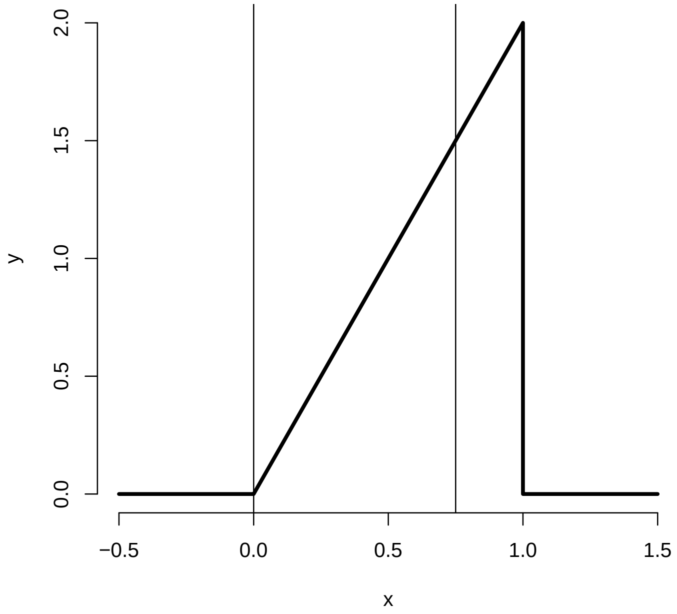
If we ask "What percentage of days do between 50% and 75% of the calls get addressed?," then the area under within the lines corresponds to the percentage of days that that range of calls get addressed.

It's roughly $0.25 \cdot 1.25 = 0.31$ or
"31% of the time (in days) between 50% and 75% of calls are addressed on a day."

Example 2:

What is the probability that 75% or fewer calls get addressed?

```
> x <- c(-0.5, 0, 1, 1, 1.5)
> y <- c(0, 0, 2, 0, 0)
> #par(pin = c(2,2))
> plot(x,y, lwd=3, frame=F, type="l")
> abline(v=0)
> abline(v=0.75)
```



The area is $0.5(0.75 \cdot 1.5) = 0.5625$.

In R, the command *pbeta()* of x will return the probability of being lower than x:

```
> pbeta(0.75, 2, 1)
```

```
[1] 0.5625
```

CDF and Survival Function

The Cumulative Distribution Function of a random variable X is defined as the function:

$$F(x) = P(X \leq x)$$

This definition applies to both discrete and continuous variables.

The Survival Function of a random variable X is defined as:

$$S(x) = P(X > x)$$

Notice that $S(X) = 1 - F(X)$.

For continuous variables the PDF is the derivative of the CDF.

Example:

What are the Survival Function and the CDF from the density considered previously?

For $1 \geq x \geq 0$
(We must exclude unanswered calls, i.e. outside of $x = [0, 1]$, for this density to make sense),

$$\begin{aligned} F(x) &= P(X \leq x) \;=\; \frac{1}{2}\, base \times height \\ &= \frac{1}{2}(x) \times (2x) \;=\; x^2 \\ S(x) &= 1 - x^2 \end{aligned}$$

In R, we can use the *pbeta()* function to get the probabilities:

```
> pbeta(c(0.4,0.5,0.6), 2, 1)
```

```
[1] 0.16 0.25 0.36
```

Quantiles

The α^{th} quantile of a distribution with the distribution function F is the point x_α so that:

$$F(x_\alpha) = \alpha$$

A percentile is simply a quantile with α expressed as a percent.

The median is the 50^{th} percentile, or the 0.5^{th} quantile, and can be expressed as $x_{0.5}$.

Consider that 0.5 of the area lies below $x_{0.5}$ and 0.5 of the area lies above it. We can express that as: $0.5 = F(x_{0.5})$.

Recall that the cumulative distribution function for our example is $F = x^2$. To find $x_{0.5}$, we must invert the function F.

$$\begin{aligned} x &= F^{-1}(x) = \sqrt{x} \\ x_{0.5} &= \sqrt{0.5} = 0.707 \end{aligned}$$

Therefore, about 70% of calls being answered on a random day is the median.

In general, the distribution function describes a point x_α along a density for which the area α lies below it, and the area $1 - \alpha$ lies above it.

In R, we can approximate the quantiles for common distributions with the function *qbeta()*:

```
> qbeta(0.5,2,1)
```

```
[1] 0.7071068
```

Probability Summary

A probability model uses assumptions to connect the sample data (which is limited) to the population. We want connect the sample median to the population median; the population median is the estimand, and the sample median is the estimator.

Expected Values - Discrete Random Variables

The Expected Value is the mean of a random variable, and is the center of its distribution.

For a discrete random variable X with PMF p(x), it is defined as:

$$E[X] = \sum_x x \cdot p(x)$$

We use the square brackets [] to denote the expected value.

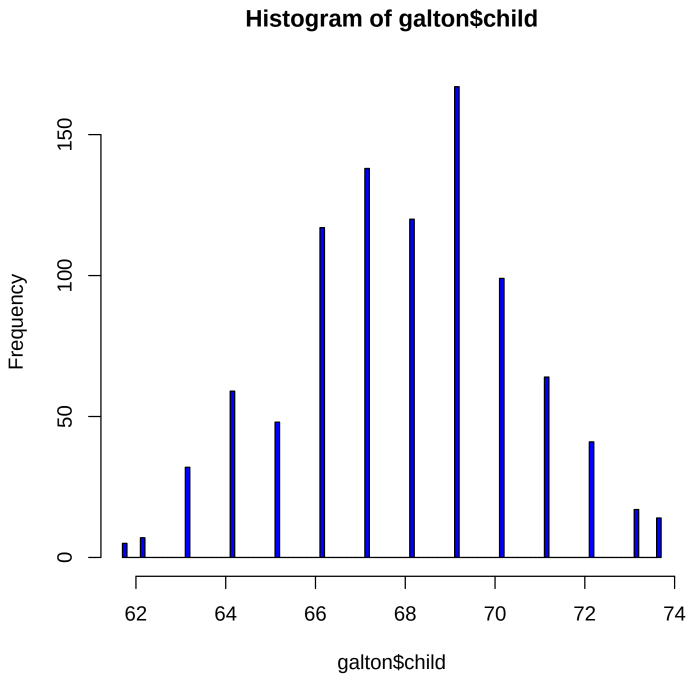
$E[X]$ represents the center of mass of a collection of locations and weights, $\{x, p(x)\}$.

R Example - Expected Value

This example uses R, the libraries ”Manipulate” and ”UsingR”, and the UsingR dataset ”Galton.

```
> library(manipulate)
> library(UsingR)
> data(galton)
> myHist <- function(mu){
+     hist(galton$child, col="blue", breaks=100)
+     lines(c(mu,mu), c(0,150), col="red", lwd=5)
+     mse <- mean((galton$child -mu)^2 )
+     text(63,150, paste("mu = ",mu ))
+     text(63,140, paste("Imbalance = ", round(mse,2) ))
+ }
> #manipulate(myHist(mu), mu=slider(62,74, step=0.5))
```

The manipulate() library makes interactive plots, so it won’t work in this .pdf document (and it’s commented out) but the histogram looks like this:



Example - expected value of a discrete variable

Suppose that a die is rolled and X is the number that is face up.
What is the expected value of X?

$$E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

Expected Value - Continuous Random Variables

For a continuous random variable X with a density f the expected value is defined as:

$$E[X] = \text{ the area under the function } tf(t)$$

This definition borrow from the definition of center of mass for a continuous body.

Example - expected value of a continuous variable

Consider a density where $f(x) = 1$ for x between 0 and 1.
Is this a valid density?
Suppose X follows this density. What is its expected value?

$f(x)$ describes a block shaped density.
It's valid because $y > 0$ everywhere that x is positive,
and it has a calculable area (base x height = 1 x 1 = 1).

The function $tf(t)$ in this case is a right triangle where $x = y$.
We can integrate that to find the area, or we can simply compute the area of a triangle: (1/2 x base x height = 0.5).
So 0.5 is the mean and the expected value, and it is also intuitively the center of mass for the block.

Rules of Expected Values

- The expected value is a linear operator.
- If a and b are not random and X and Y are two random variables, then:

  $E[aX + b] = aE[X] + b$

  $E[X + Y] = E[X] + E[Y]$

- but:

  $E[X \cdot Y] \neq E[X] \cdot E[Y]$

  $E[X^2] \neq E[X]^2$

examples - rules of expected values

You flip a coin X and simulate a uniform random number Y.
What is the expected value of their sum?

The random uniform density is the block-like density which yields numbers between 0 and 1; the expected value of the uniform density is 0.5.
The coin flip yields only 0 or 1, and its expected value is also 0.5.

$$E[X+Y] = E[X] + E[Y] = 0.5 + 0.5 = 1$$

You roll a die twice. What is the expected value of the average?

Let X_1 and X_2 be the results of two dice rolls. The average of course is their sum divided by their count.

$$E[(X_1 + X_2)/2] = \frac{1}{2}(E[X_1] + E[X_2]) = \frac{1}{2}(3.5 + 3.5) = 3.5$$

That case can be generalized:
Let X_i for $i = 1, ..., n$ be a collection of random variables, each from a distribution with mean μ.
Then the expected value of the sample average of X_i is found by:

$$\begin{aligned} E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] &= \frac{1}{n}E\left[\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu = \mu \end{aligned}$$

So if all the variables have the same mean, then their average also has that mean.
Therefore, the expected value of the sample average is the population mean as well; it's exactly what we would like it to be.
When the expected value of an estimator is what it is trying to estimate, we say that the estimator is unbiased.

Variance

The variance of a random variable is a measure of spread.

If X is a random variable with mean μ, the variance is defined as:

$$Var(X) = E[(X-\mu)^2]$$

the expected (squared) distance from the mean.

Densities with a higher variance are more spread out then densities with lower variance.

Computational form:

$$Var(X) = E[X^2] - E[X]^2$$

If a is constant then $Var(aX) = a^2 Var(X)$.
If you pull a constant out of a variance, the constant gets squared.

The square root of the variance is the standard deviation.
The standard deviation has the same units as X.
If you pull a constant out of a standard deviation, it doesn't get squared.

Example-variance

What's the sample variance from the result of a toss of a die?

Recall that the expected variable of a die is $E[X] = 3.5$.

$$\begin{aligned} E[X^2] &= 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6} = 15.17 \\ Var(X) &= E[X^2] - E[X]^2 \\ &= 15.17 - 3.5^2 = 2.92 \end{aligned}$$

Interpreting Variances

Chebyshev's inequality is useful for interpreting variances:
The probability P that a random variable X is more than or equal to k standard deviations σ from its mean μ is less than or equal to $1/k^2$ for all distributions.

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

In other words, the probability that a random variable lies beyond k standard deviations from its mean is less than $1/k^2$.

$$\begin{aligned} 2\sigma &\rightarrow 25\% \\ 3\sigma &\rightarrow 11.11\% \\ 4\sigma &\rightarrow 6.25\% \\ 7\sigma &\rightarrow 2.04\% \\ 10\sigma &\rightarrow 1\% \end{aligned}$$

These are upper bounds and the actual probability might be much smaller. With the normal distribution, for example, the probability that a variable lies 3 standard deviations away or further is 1%.

example 1 - interpreting variance

IQs are often said to be normally distributed with a mean of 100 and a sd of 15. What is the probability of a randomly drawn person having an IQ higher than 160 or below 40?

- Thus we want to know the probability of a person being more than 4 standard deviations from the mean.
- Thus Chebyshev's inequality suggests that this will be no larger than 6%.
- IQs distributions are often cited as being bell shaped, in which case this bound is very conservative.
- The probability of a random draw from a bell curve being 4 standard deviations from the mean is on the order of 10^{-5} (one thousandth of one percent).

example 2 - interpreting variance

A former buzz phrase in industrial quality control is Motorola's "six sigma", whereby businesses are suggested to control extreme events or rare defective parts.

Chebyshev's inequality states that the probability of a "Six Sigma" event is less than $1/6^2 = 2.78\%$.

If a bell curve is assumed, then the probability is on the order of 10^{-9} or one ten millionth of a percent.

Moments

Variances and Means are both types of Moments of a distribution, and they are the first two moments, and the most important moments of a distribution.

Independence

Independent Events

Two events A and B are independent if

$$P(A \cap B) = P(A)P(B)$$

Two random variables X and Y are independent if for any two sets A and B

$$P([X \in A] \cap [Y \in B]) = P([X \in A])P([Y \in B])$$

If A is independent of B then:

- A^c is independent of B.
- A is independent of B^c.
- A^c is independent of B^c.

example - independent events

What is the probability of getting two consecutive heads?

- $A = \{\text{Head on flip 1}\} \approx P(A) = 0.5$
- $B = \{\text{Head on flip 2}\} \approx P(B) = 0.5$
- $A \cap B = \{\text{Head on flip 1 and 2}\}$
- $A \cap B = P(A)P(B) = 0.5 \cdot 0.5 = 0.25$

The multiplication of probabilities is only valid for the intersection of independent events. If for any reason, known or unknown, the events are dependent then muliplying the probabilities isn't appropriate.

An example is the Dr. Meadow SIDS testimony in which a mother was convicted of murder after two children died of SIDS. Dr. Meadow calculated the probability of SIDS occurring twice in the same household as if they were independent events, however SIDS could have a genetic or environmental component which would mean that the events are not independent.

Thus, $P(A_1 \cap A_2)$ is not necessarily equal to $P(A_1)P(A_2)$.

Joint Distribution

If a collection of random variables $X_1, X_2, ..., X_n$ are independent, then their joint distribution is the product of their individual densities or mass functions. That is, if f_i is the density for random variable X_i, we have that

$$f(x_1, ..., x_n) = \prod_{i=1}^{n} f_i(x_i)$$

IID Random Variables

Independent and identically distributed random variables. Identically distributed means that they come from the same distribution. Coin flips are IID.
iid random variables are the default model for random samples. We will assume our data are iid if we assume it is random even if it hasn't been sampled randomly.
Most of the important theories of statistics are founded on the assumption that variables are iid.

example - iid

Suppose we flip a biased coin with success probability p n times, what is the joint density of the collection of outcomes?

These random variables are iid with densities $p^{x_i}(1-p)^{1-x_i}$, therefore:

$$\begin{aligned} f(x_1, ..., x_n) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ &= p^{\sum x_i}(1-p)^{n-\sum x_i} \end{aligned}$$

(To be clear, we're putting the summations in the exponents.)

That implies that the order of the sequence isn't important, i.e. p(1,1,0,0) is the same as p(0,1,0,1). The probability depends only on the total number of successes and failures.

Correlation

The covariance between two random variables X and Y is defined as

$$Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)] = E[XY] - E[X]E[Y]$$

where μ is the mean.
Covariance is a measure of how two variables are linearly unrelated.

1. $Cov(X, Y) = Cov(Y, X)$
2. $Cov(X, Y)$ can be positive or negative.
3. $|Cov(X, Y)| \leq \sqrt{Var(X)Var(Y)}$

   where $\sqrt{Var(X)Var(Y)} = sd(X)sd(Y)$

That third statement implies correlation:

$$Cor(X, Y) = Cov(X, Y)/\sqrt{Var(X)Var(Y)}$$

which has the following properties:

1. $-1 \leq Cor(X, Y) \leq 1$
2. $Cor(X, Y) = \pm 1$ if and only if $X = a + bY$ for some constants a and b.
3. $Cor(X, Y)$ is unitless.

   sd has units of X or Y.
4. X and Y are uncorrelated if $Cor(X, Y) = 0$

   They are also uncorrelated if $Cov(X, Y) = 0$.
5. X and Y are more positively correlated the closer $Cor(X, Y)$ is to 1.
6. X and Y are more negatively correlated the closer $Cor(X, Y)$ is to -1.

These properties describe the population correlation, which is a statement about the joint population density of the random variables X and Y. The sample correlation is an estimate of the population correlation.

Let $X_{i i=1}^{n}$ be a collection of random variables:

- When the X_i are uncorrelated:

  $$Var\left(\sum_{i=1}^{n} a_i X_i + b\right) = \sum_{i=1}^{n} a_i^2 Var(X_i)$$

  The constant a is pulled out and squared.

  The constant b is irrelevant because shifting things doesn't affect variance.

- A commonly used subcase from this property is that if a collection of random variables X_i are uncorrelated, then the variance of the sum is the sum of the variances:

  $$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i)$$

  X_i must be independent. If they aren't indepedent, then that property isn't true.

- Therefore, it is sums of variances that tend to be useful, not sums of standard deviations. In other words, the standard deviation of the sum of a bunch of independent random variables is the square root of the sum of the variances, not the sum of the standard deviations.

Sample Mean $\bar{X}$

Suppose X_i are iid with variance σ^2:

$$\begin{aligned} Var(\bar{X}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var\left(X_i\right) \\ &= \frac{1}{n^2} n\sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

where $\bar{X}$ is the Sample Mean of the random variable's values.

Suppose we had 10 dice and rolled them and took the average $\bar{X}$. $\bar{X}$ is a random number and it has a probability mass function: it can be an integer from 1 to 6, and will likely be a bell-shaped curve. We know the center of the distribution is at 3.5 for a six-sided die. We now know that the sample mean variance has to be the variance σ^2 of a single die divided by the number of dice n. And if we take the square root of the sample mean variance, we get the standard error $\sigma/\sqrt{n}$

- $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$ when X_i are independent with common variance.
- $\frac{\sigma}{\sqrt{n}}$ is the standard error of the sample mean.
- The standard error of the sample mean is the standard deviation of the distribution of the sample mean.
- σ is the standard deviation of the distribution of a single observation.
- The sample mean has to be less variable than a single observation, therefore it's reasonable that its standard deviation is divided by some amount related to the number of observations; that amount is $\sqrt{n}$.
- Both the standard deviation and the standard error have units of standard deviation. Standard error is a measure of the variability of the average of observations. Standard deviation is a measure of the variability of individual observations.

Sample Variance s^2

Sample variance is the sample average squared deviation from the empirical mean:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

- The sample variance is an estimator of variance σ^2.
- Here is a version that's quicker for calculation:

  $$s^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1}$$

  which is analagous to $E[X^2] - E[X]^2$ that approximates the variance for the population calculation.
- The sample variance is (nearly, except for the -1) the mean of the squared deviations from the mean.

The sample variance is unbiased $n-1$

To explain why we use $n-1$ instead of n in the denominator, let's look at the expected value of the numerator. First, recall the formula for the variance of a random variable:

$$Var(X) = E[X^2] - E[X]^2$$

Thus:

$$Var(X) + E[X]^2 = E[X^2]$$

And recall that the mean is μ. Therefore if we want to calculate the expected value of the sample mean squared $E[\bar{X}^2]$, we can plug in the expected value of the mean squared μ^2 plus the variance of the mean $Var(\bar{X})$. Likewise, we can calculate $E[X_i^2]$ by substituting it with μ^2 and adding $Var(X_i)$.

$$\begin{aligned}
E\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right] &= \sum_{i=1}^{n} E[X_i^2] - nE[\bar{X}^2] \\
&= \sum_{i=1}^{n} \left\{Var(X_i) + \mu^2\right\} - n\left\{Var(\bar{X}) + \mu^2\right\} \\
&= \sum_{i=1}^{n} \left\{\sigma^2 + \mu^2\right\} - n\left\{\sigma^2/n + \mu^2\right\} \\
&= n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2 \\
&= (n-1)\sigma^2
\end{aligned}$$

I got lost in there, but basically the $n-1$ term is the correct denominator to avoid a bias in the sample mean.

Some points to avoid confusion.

- Suppose X_i are iid with mean μ and variance σ^2.
- s^2 estimates σ^2.
- The calculation of S^2 involves dividing by $n-1$.
- $s/\sqrt{n}$ estimates the standard error of the mean $\sigma/\sqrt{n}$.
- s^2/n estimates the variance of the mean σ^2/n
- $s/\sqrt{n}$ is called the sample standard error (of the mean).
- Don't confuse that $/\sqrt{n}$ with the $/(n-1)$ which is only for calculating s^2.

R Example - Sample Variance

This example uses R, the libraries ”Manipulate” and ”UsingR”, and the UsingR dataset ”father.son”.

```
> library(UsingR)
> data(father.son)
> x <- father.son$sheight
> n <- length(x)
> hist(father.son$sheight,n,breaks=10,col="skyblue")
> round(c(sum((x-mean(x))^2)/(n-1), var(x), var(x)/n,
+         sd(x), sd(x)/sqrt(n)),2)
```

```
[1] 7.92 7.92 0.01 2.81 0.09
```

where we have a list of:

1. s^2, manual calculation of sample variance of heights (inches^2).
2. R calculation of s^2.
3. Variance of the mean of heights (inches^2).
4. Standard deviation of heights (inches).
5. Sample standard error of the mean of heights (inches).