MATH 324: Statistics

Julian Lore

Last updated: January 18, 2018

Notes from Masoud Asgharian's Winter 2018 lectures.

Contents

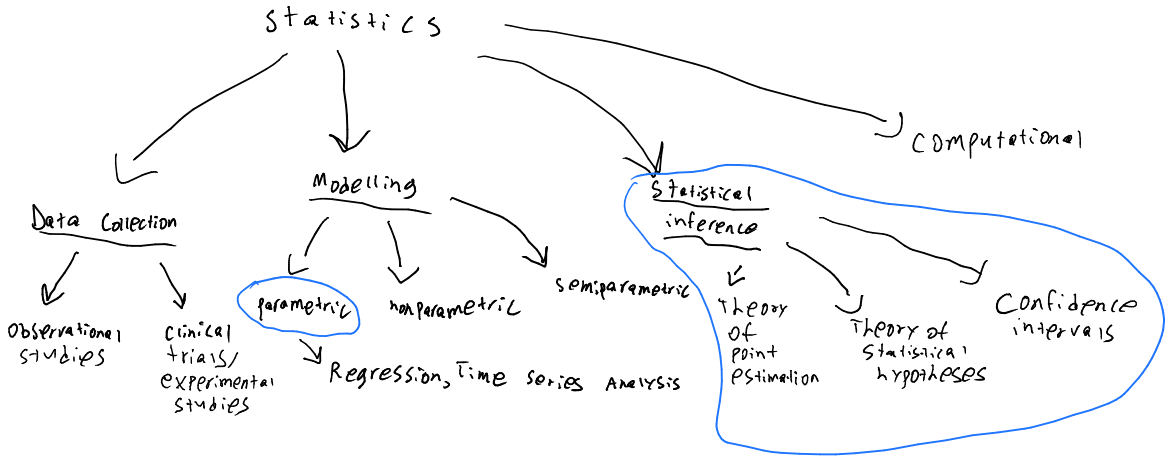
1 01/09/18

What we will cover this semester Will essentially cover chapter 8,9,10. For chapter 11, he will give us his own notes. The first 6 sections of chapter 13 and a few sections from chapter 14. Occasionally we will go back to chapter 7 to revisit things like the t distribution. In 323, we made probabilistic models. Statistics is the breach in which we connect these models to real life. Otherwise, those are just models. A core part of data analysis and data sciences is statistics and computer science.

1.1 Overview - What is Statistics?

Inductive logic, we have a sample from the population we want to make inference about. With this data, we want to extend the results to the whole population. From small to large,

sample to the population.



- Observational studies: we go to the population and make observations.
- Experimental studies: give test subjects something, i.e. give them cigarettes when trying to test for if cigarettes lead to cancer. Need to account for causation, other factors that can affect outcome. In order to do so we have to keep their diet and other factors controlled. We must also have some sort of randomization, we can't send all males to one group and all females to another, as males may have a tendency to smoke or something of the like. These are also called clinical trials.
- When we have data, the next step is modeling. May occasionally speak of this, but this is not part of the course. There are different approaches to modeling, can be split into 3 parts.
  - Parametric: the salary is distributed like a distribution (ex. Gamma), but we don't know the parameters. Take for example, we always know that the normal distribution is a bell curve, but we don't know where it's centered. Very useful, but we might have a miss-specification. How do we know our models are correct? Most of the time we will be talking about **parametric** models.
  - Semiparametric
  - Nonparametric: since we don't know if parametric models are correct, we make no assumption about the distribution. We just assume that $X \sim F$, all we assume about F is that it's continuous, nothing more. This is an infinite dimensional vector. Why? How do we know a function? We have a vector for F, like $F(1), F(2, \ldots)$. How do we approximate this? $X_i \overset{iid}{\sim} F, i = 1, 2, \ldots, n$. n

patients, with all the same distribution. So $F(t) = P(X \leq t)$. What does this tell us? The proportion of time that x falls below t. So with n samples, how do we mimic this? We count the number of observations below t, i.e. $\frac{\#X_i \leq t}{n}$, which is an approximation of the above. This is an empirical observation. More mathematically:

$$\varepsilon(t) = \begin{cases} 1 & \text{if } t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

So we have $\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon(t - x_i)$. This gives us a binomial distribution. But we are assuming they are all the same distribution.

Nonparametric approaches are good for functions of single variables, but not for multi variables, which is what semiparametric was made for.

- Bayesian inference: when you learn that $X \sim N(\mu, \sigma^2)$, X is normally distributed and μ is the average of the whole population. Bayes' approach says that these parameters are not constants, these are random variables themselves. Bayes did not look at probability as a frequentest approach, not the proportion of when something arrives (frequentest approach works when we have a huge sample). The other approach that Bayes had was an updating approach, that our parameters are unknown. This is good for when you have a stream of data (machine learning is a prime example). We have a lack of knowledge and then we update it using Bayesian's approach. $\rightarrow X|\mu, \sigma^2 \sim N(\mu, \sigma^2)$, i.e. the parameters are also normally distributed.

Most of the time we'll be at parametric modeling and statistical inference.

1.2 Point Estimation

What do we mean by point estimation? A scientific guess about the unknown parameter of the population. Consider the following situation:
$x_1, \ldots, x_n \sim N(\mu, 1)$ (usually interested in the normal distribution, binomial and poisson). Suppose this is the IQ of high school graduates in Canada (the X_i are numbers). Why do we call this distribution normal? Because for a healthy population, most of the weight should be in the middle, just like the bell curve. The Normal distribution is especially important for modeling error. For insurance companies, we see at the tails that there aren't many large claims.
We want to find μ. Recall that $E(X_i) = \mu, i = 1, 2, \ldots, n$ (if they all have the same observations, they have the same mean).
First, what is a point estimation? What properties should it have? If we know the value of

μ, we have the whole thing, can calculate everything. How do we estimate this? The whole population is huge, so we take a sample part of the population, mimicking the real μ, getting $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. $\overline{X}_n$ is useful, but $\overline{X}_n - \mu$ isn't, as there's an unknown we have here.

Statistic A function of observations that does not depend on any unknown parameter.

Ex $\overline{X}_n$ is a statistic. $\overline{X}_n - \mu$ is not.

Estimator A statistic that aims at estimating an unknown parameter (we want to work with it). For example, if μ moves from $-\infty$ to ∞, we want to have an estimator that also has the same range, not one that is strictly positive. Example: $\overline{X}_n$ is an estimator. However, consider:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

This is a statistic, but not an estimator, it always returns a positive value. Also, take for example in physics, where each measure has a unit of measurement. This statistic wouldn't even be the same unit, so it once again is a bad estimator.
When we take a mean and try to estimate it, the next step is to figure out how we quantify possible bias.

$$\varepsilon = |\overline{X}_n - \mu|$$

We can use Tchbycshev's inequality to put a bound on the error.

$$P(|X - \underbrace{E(X)}_{\mu_x}| > k\sqrt{\underbrace{Var(X)}_{\sigma_x^2}})$$
$$P(|X - \mu_x > k\sigma_x) \leq \frac{1}{k^2}$$

Very useful, assume very little but get lots of information. One of the big hammers of probability and statistics. The only thing we assume here is the existence of the second moment.
Consider $k = 3$.

$$P(|X - \mu_x| > 3\sigma_x) \leq \frac{1}{9}$$
$$P(|X - \mu_x| \leq 3\sigma_x) \geq 1 - \frac{1}{9} \approx \%89$$

Without knowing anything else about the distribution, this tells us that about 89% of the population is within 3 times the variance of the mean.

2 01/11/18

Last lecture we learned about statistics, estimators and how we can measure deviation from the target and the estimation.
We had n random variables: $x_1, \ldots, x_n, \mu \to$ IQ in the population. We want to have a scientific guess of the average IQ (there are many more examples, like salary). Our n random variables are n random people chosen.
We then arrive at $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, a scientific guess mimicking μ but in the sample population.
How much deviation do we have? $|\overline{X}_n - \mu|$
Since our observations are random, then $\overline{X}_n$ will also be random, i.e. $\overline{X}_n$ is a random variable itself. Each person gives us a different deviation, so we need a way to summarize all of this information, say, the expected value.

$$E[|\overline{X}_n - \mu|]$$

Another way to summarize it is with probability.

$$P(|\overline{X}_n - \mu| > \varepsilon)$$

What is the chance that what we produce is not within ε of the target? Often times we want to bound these things. What do you think might happen if instead of taking a sample of $n = 50$, we take $n = 100$? As n increases, we should get closer to the target. But, the more samples I take, the more it'll cost me. So we want to have a balance. We want the distance from the target to be within some sort of value.

$$P(|\overline{X}_n - \mu| > \varepsilon) \leq \delta$$

Take for example a spam filter. Something is either spam or not spam. We start with messages and then start checking. We want to know how many messages we should check, i.e. how big our training set should be.
We will use Tchbycshev's Inequality for this!

2.1 Chebyshev/Tchbycshev's Inequality

Let X be a random variable. Suppose $h(x)$ is a positive function (i.e. the range of this function consists of positive values). We can show that

$$P(h(x) \geq \lambda) \leq \frac{E[h(x)]}{\lambda}$$

for any $\lambda > 0$, if $E[h(X)] \leq \infty$, i.e. it exists. This is called **Markov's Inequality**
When we say that the expected value of a random variable exists, we mean $E[|X|] < \infty$. When we talk about existence of a moment, we check the absolute value, but the actual value does not have an absolute value, it is just $E[X]$. Why? The trouble is when X can take positive and negative values and is not bound.

$$E[X] = \sum_{i=1}^{\infty} X_i P(X = x_i)$$

What if we have infinite values that we can take?

Recall from Calculus $\sum_{n=1}^{\infty} \frac{(-1)^n}{n} < \infty$ is convergent, but not absolutely convergent because $\sum_{n=1}^{\infty} \left|\frac{(-1)^n}{n}\right| = \infty$. Riemann has a result such that if a series is convergent but not absolutely convergent (like the example just mentioned), then it can converge to any real number (if we reorder the terms). Thus we don't like this and must check for absolute convergence for moments, or else the expected value will depend on the order we consider the numbers in.
Recall the theorem that says if we have a function of a random variable, we don't need its distribution, we can directly use the distribution of X. Note that integrals are another form of sums, we can use similar notation with x as a subscript to denote ranging over all x.

$$E[h(x)] = \int_x h(x) f_x(x)\ dx = \left(\int_{x:h(x)\geq\lambda} + \int_{x:h(x)<\lambda} h(x) f_x(x)\ dx \right)$$

(Note that the two integrals both apply on the right side)

$$\geq \int_{x:h(x)\geq\lambda} h(x) f_x(x)\ dx \geq \lambda \int_{x:h(x)\geq\lambda} f_x(x)\ dx = \lambda P(h(x) \geq \lambda)$$

So what did we get?

$$E[h(x)] \geq \lambda P(h(x) \geq \lambda)$$

$$P(h(x) \geq \lambda) \leq \frac{E[h(x)]}{\lambda}$$

Now consider: $h(x) = (x-\mu)^2$. Then what do we have?

$$P(|x-\mu| \geq \lambda) = P([x-\mu]^2 \geq \lambda) \overset{\text{By Markov's Inequality}}{\leq} \frac{E[(x-\mu)^2]}{\lambda^2}$$
$$P(|x-\mu| \geq \lambda) \leq \frac{Var(x)}{\lambda^2}$$

Replace λ by $k\sigma_x$ where $\sigma_x = \sqrt{Var(x)}$

k is a constant, so we get:

$$P(|x-\mu| \geq k\sigma_x) \leq \frac{Var(x)}{k^2\sigma_x^2} = \frac{Var(x)}{k^2 Var(x)} = \frac{1}{k^2}$$

This is **Tchbycshev's Inequality**. For $k = 3$ we have:

$$P(|x-\mu| \geq 3\sigma_x) \leq \frac{1}{9}$$
$$P(|x-\mu| \leq 3\sigma_x) \geq \frac{8}{9} \approx 88\%$$

What does this say? For any random variable with 2 moments, 88% of the values fall within $3\sigma_x$s from the center of gravity (mean). This is a very crude lower bound that required almost no assumptions, all we need is that $\mu = E(x)$ and $\sigma_x^2 = Var(x)$ and the existence of the second moment.
Back to where we were before with $|\overline{X}_n - \mu|$:

$$P(|\overline{X}_n - \mu| > 3\sigma_{\overline{X}_n}) \leq \frac{1}{9}$$

Note that here, it must be true that $\mu = E(\overline{X}_n)$. Is this true?

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$E[\overline{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$$

Recall: $E[cY] = cE[Y]$, think of expected values like integrals and sums, they have the same properties.

$$= \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i]$$

Remember that $X_1, \ldots, X_n \sim F$, i.e. they all have the same distribution! So $E[X_i] = \mu, i = 1, 2, \ldots, n$

$$= \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

Now how do we use this in practice? If everyone is sent to the population and asked to take a sample of size 10 (the same size for everyone) and everyone makes their own $\overline{X}_n$, their own sample average and then we take the average of all the sample averages and we obtain the actual average of the population, i.e. this is an average of averages (this is difficult though, as we need all the possible averages of size 10; in practice we only use one sample, more on this later).

Example Suppose we have the following 0-1 random variable representing what people will vote for

$$X_i = \begin{cases} 1 & \text{if NDP} \\ 0 & \text{otherwise} \end{cases}$$

We know that $X_i \sim \text{Bernoulli}(p)$, $p = P(X_i = 1)$

$$X_1, \ldots, X_n \sim p = (P(X_i = 1), i = 1, 2, \ldots, n)$$

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \hat{p}_n$$

Side Note about Picking with Replacement

$$P(X_2 = 1 | X_1 = 1) = \frac{M-1}{N-1}$$

(after taking 1 in favor of NDP, where N is total population size and M is total number of people in favor of NDP) Using Total Probability Theorem we get:

$$\begin{aligned} P(X_2 = 1) &= P(X_2 = 1|X_1 = 1)P(X_1 = 1) + P(X_2 = 1|X_1 = 0)P(X_1 = 0) \\ &= \frac{M-1}{N-1} \cdot \frac{M}{N} + \frac{M}{N-1}\left(1 - \frac{M}{N}\right) = \frac{M}{N} \\ P(X_2 = 1) &= \frac{M}{N} \\ P(X_2 = 1|X_1 = 1) &= \frac{M-1}{N-1} \end{aligned}$$

These are identically distributed, but not independent, this replacement is what differs hypergeometric from binomial. But when the sample size is very large we can just use binomial (also we don't ask someone who they're voting for twice).

Note we just showed that

$$P(|\overline{X}_n - \mu| > k\sigma_{\overline{X}_n}) \leq \frac{1}{k^2}$$

Recall that if $X_i \sim$ Bernouilli(p), then $E[X_i] = p$

$$P(|\overline{X}_n - \mu| \geq k\sigma_{\overline{X}_n}) = P(|\hat{p}_n - p| > k\sigma_{\hat{p}_n})$$

Recall that:

$$\begin{aligned} \sigma^2_{\overline{X}_n} &= Var(\overline{X}_n) \\ Var(cY) &= c^2 Var(Y) \\ Var(X \pm Y) &= Var(X) + Var(Y) \pm 2Cov(X, Y) \end{aligned}$$

If $Cov(X, Y) = 0$, $Var(X \pm Y) = Var(X) + Var(Y)$. If X and Y are independent, $Cov(X, Y) = 0$.

$$\begin{aligned} Cov(X, Y) &= E(XY) - E(X)E(Y) \\ X \perp\!\!\!\perp Y \implies E[g(x)h(y)] &= E[g(x)]E[h(y)] \\ X \perp\!\!\!\perp Y \implies E(XY) &= E(X)E(Y) \text{ therefore } Cov(X, Y) = 0 \end{aligned}$$

Thus,

$$Var(\overline{X}_n) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}Var(\sum_{i=1}^{n} X_i)$$

So,

$$\frac{1}{n^2}Var\left(\sum_{i=1}^{n} X_i\right) \overset{\text{Assuming that } X_i\text{s are ind}}{=} \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$$

Remember that $X_i \sim F$

$$\overset{\text{Assuming that } X_i\text{s have the same variance}}{=} \frac{1}{n^2} \cdot nVar(X_i) = \frac{Var(x)}{n}$$

Thus we have:

$$Var(\overline{X}_n) = \frac{Var(X)}{n}$$

i.e. variance gets smaller and smaller as the population size increases, so $\overline{X}_n$ gets closer and closer to its center.

$$Var(\hat{p}_n) = Var(\overline{X}_n) = \frac{Var(X)}{n}$$

Where $X \sim$ Bernouilli(p), so $Var(X) = p(1-p)^n$

$$Var(\hat{p}_n) = \frac{p(1-p)}{n}$$

$$P\left(|\hat{p}_n - p| > k\sqrt{\frac{p(1-p)}{n}}\right) \leq \frac{1}{k^2}$$

Now, back to the original problem:

$$P(|\hat{p}_n - p| > \varepsilon) \leq \delta$$

where ε, δ are known (and small values). Next class we will see how to choose values to satisfy this.

3 01/16/18

Recall Last class we were talking about voting. We got to the point of estimating thousands of votes. We want to estimate the amount of Canadians voting NDP.

$$X_i = \begin{cases} 1 & \text{NDP} \\ 0 & \text{otherwise} \end{cases}$$

$i = 1, \ldots, n$

$p(X_i = 1) = p, i = 1, \ldots, n$, we want to estimate this. Why can we make the assumption that this is p? If we chose someone randomly, the probability that one of them is voting NDP is p, the same for each sample. We try not to get samples from the same family, so that the variables remain independent, as they might have the same views. We use the following notation to signify independence:

$$\perp\!\!\!\perp_{i=1}^{n} X_i$$

$$\hat{p}_n = \frac{1}{n} \underbrace{\sum_{i=1}^{n} x_i}_{\substack{\text{Number of people} \\ \text{in sample voting NDP}}} = \overline{X}_n$$

$$\varepsilon = |\hat{p}_n - p|$$

$$p(|\hat{p}_n - p| > \overbrace{\delta}^{\text{given}}) \leq \frac{E[|\hat{p}_n - p]^2}{\delta^2}$$

$$E[|\hat{p}_n - p|^2]$$

$$E(\hat{p}_n) = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

$$E(X_i) = 1 \cdot p(X_i = 1) + 0 \cdot p(X_i = 0) = 1 \cdot p + 0 \cdot (1 - p) = p(\text{Expected value of a Bernoulli variable})$$

$$\implies E(\hat{p}_n) = \frac{1}{n}\sum_{i=1}^{n} p = \frac{1}{n}np = p$$

Expectation of variable minus its expectation squared is variance:

$$\begin{aligned} E[|\hat{p}_n - p|^2] &= Var(\hat{p}_n) \\ Var(\hat{p}_n) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \left(\frac{1}{n}\right)^2 Var\left(\sum_{i=1}^{n} X_i\right) \\ Var\left(\sum_{i=1}^{n} X_i\right) &\overset{\text{Thm 5.12(b), p271}}{=} \sum_{i=1}^{n} Var(X_i) + 2\sum\sum_{i\le i<j\le n} Cov(X_i, X_j) \end{aligned}$$

Recall that $\perp\!\!\!\perp_{i=1}^{n} X_i \implies Cov(X_i, X_j) = 0, \forall i \neq j$, so all the terms in the double sum will be 0.

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

Now we must calculate the variance of X_i. We already have the first moment:

$$\begin{aligned} Var(X_i) &= E(X_i^2) - [\underbrace{E(X_i)}_{p}]^2 \\ E(X_i^2) &= 1^2 \cdot p(X_i = 1) + 0^2 \cdot p(X_i = 0) \\ &= 1 \cdot p + 0 \cdot (1 - p) = p \end{aligned}$$

$$\begin{aligned} Var(X_i) &= p - [p]^2 = p - p^2 = p(1-p) \\ Var\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} p(1-p) = np(1-p) \\ Var(\hat{p}_n) &= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n} \end{aligned}$$

Now, using Chebyshev's inequality:

$$P(|\hat{p}_n - p| > \delta) \le \frac{E[|\hat{p}_n - p|^2]}{\delta^2} = \frac{Var(\hat{p}_n)}{\delta^2}$$

$$P(|\hat{p}_n - p| > \delta) \leq \frac{E[|\hat{p}_n - p|^2]}{\delta^2} = \frac{p(1-p)}{n\delta^2}$$

Is this useful? As n tends to ∞, the bound goes to 0, meaning, no matter what δ you choose here, the chance gets smaller and smaller, i.e. the bigger n gets, the more were on track, heading to the right direction. But we don't know p, so how is this useful? But we know something else:

$$p(1-p) \leq \frac{1}{4}$$

Why?

$$\begin{aligned} f(x) &= x(1-x) \\ \frac{df}{dx} &= 1 - 2x = 0 \implies x = \frac{1}{2} \\ \frac{d^2 f}{dx^2} &= -2 \end{aligned}$$

Thus we get a concave function with $f\left(\frac{1}{2}\right)$ as the max.

$$f\left(\frac{1}{2}\right) = \frac{1}{2}\left(1 - \frac{1}{2}\right) = \frac{1}{4}$$

$$p(|\hat{p}_n - p| > \delta) \leq \frac{1}{4n\delta^2}$$

So now for any δ we can determine how far off our estimate will be. Another application is sample size determination with $\delta = 0.01$, we want the deviation from the target to be less than 1% and we don't want it to fail (from being in the range) more than 5% of the time.

$$\begin{aligned} \frac{1}{4n\delta^2} &= 0.05 \\ n &= \frac{1}{4(0.05)(0.01)^2} \end{aligned}$$

This is conservative, as we've put a bound without knowing p and now we know how many samples we should obtain.

The bounds like Chebyshev's are very crude inequalities with no assumptions, when we have

things like binary variables, we can make more assumptions and get better bounds, won't need to take as many samples. If we want to train a system (i.e. machine learning or spam filtering), if it's not expensive we don't care as much, but in other cases like sampling patients, we might want to get better bounds so we can keep the cost lower.

$$p(|\overline{X}_n - \mu| > \delta)$$

We might want to minimize the distance between $\overline{X}_n$ and μ. In Euclidean space, we square the difference for the distance.

$$E[|\overline{X}_n - \mu|^2] = MSE(\overline{X}_n)(\text{Mean Squared Error})$$

Is it possible to decrease this error to 0? Aside from n being very large (i.e. everyone in the population; a census). It can be 0 if $\mu = E(X)$, meaning that $Var(X) = 0$, i.e. it is the same everywhere.

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

Now let's look at MSE more.

$$MSE(\hat{\theta}_n) = E[(\hat{\theta}_n - \theta)^2]$$

Where MSE is the estimator and θ is the estimand.

$$\begin{aligned}
&= E[\{(\hat{\theta}_n - E(\hat{\theta}_n)) + E(\hat{\theta}_n - \theta)\}^2] \\
&= E[(\hat{\theta}_n)^2 + (E(\hat{\theta}_n - \theta))^2 + 2(\hat{\theta}_n - E(\hat{\theta}_n - E(\hat{\theta}_n)))(E(\hat{\theta}_n - \theta))] \\
&= E[(\hat{\theta}_n - E(\hat{\theta}_n)^2] + E[(E(\hat{\theta}_n) - \theta)^2] + 2E[(\hat{\theta}_n - E(\hat{\theta}_n))(\overbrace{E(\hat{\theta}_n)}^{\text{constant}} - \overbrace{\theta}^{\text{constant}})] \\
E(X - \mu_x) &= 0 \\
&= Var(\hat{\theta}_n) + [\underbrace{E(\hat{\theta}_n - \theta)}_{Bias(\hat{\theta}_n)}]^2
\end{aligned}$$

Now what does Variance and Bias measure? Variance measures the reliability, the larger the variance, the less reliable our data is. If someone does the same type of experiment (in another country, with the same exact procedure) as me and comes up with another estimated value for the estimand, then this mean we don't have much credibility. We want to minimize variance.

What about bias? If $Bias(\hat{\theta}_n) = 0 \implies \hat{\theta}_n$ is an unbiased estimator. If the average of the estimator is equal to the target, then we say the estimator is an unbiased estimator. Unbiased estimators aren't a big topic but **bias is bad**. Say we want to know the salary of all Canadians. If we see that every time we take a sample and we get it over the target or under the target, then we see that we are over/underestimating, so we need to make sure the bias is (if possible) 0. But sometimes we can allow for a bit of bias but the variance goes down and makes the MSE go down dramatically as a whole. Often we work with estimators that are biased, because if we don't allow for a bit of bias, then the variance remains very high and so does the MSE.

Unbiased estimator $\hat{\theta}_n$ is said to be an unbiased estimator of θ if $E(\hat{\theta}_n) = \theta$.

Example $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, $i = 1, \ldots, n$ with μ and σ^2 both unknown.

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \implies E(\overline{X_n}) = \mu$$

$$Var(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n} Var(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$MSE(\overline{X}_n) = Var(\overline{X}_n) + [Bias(\overline{X}_n)]^2 = \frac{\sigma^2}{n} + 0^2 = \frac{\sigma^2}{n}$$

i.e. for very large n our results are unbiased.

Can we extend this example?

Suppose $X_i, X_2, \ldots, X_n$ have the same mean μ. Then

$$E(\overline{X}_n) = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \overbrace{E(X_i)}^{\mu} = \mu$$

We don't need normalized variables, just need the same distribution with the same mean, very little assumptions yet we can tell that it is unbiased.
Suppose further that $cov(X_i, X_j) = 0, i \neq j$

Then

$$\begin{aligned} Var(\overline{X}_n) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\left\{\sum_{i=1}^{n} Var(X_i) + 2\sum\sum_{1\le i\le j\le n} Cov(X_i, X_j)\right\} \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \end{aligned}$$

Assume that $Var(X_i) = \sigma^2, i = 1, \ldots, n$

$$\text{Then } Var(\overline{X}_n) = \frac{\sigma^2}{n}$$
$$\text{Then } MSE(\overline{X}_n) = \frac{\sigma^2}{n}$$

where $\sigma^2 = Var(X_i), i = 1, \ldots, n$. i.e. if $X_i's$ have the same mean μ & Variance σ^2 and $Cov(X_i, X_j) = 0, i \neq j$. Very minimal assumptions, yet we know what MSE is.

This is called **Stein's paradox**.

$$X \sim N(\mu_X, 1), i = 1, \ldots, n \to \overline{X}_n$$
$$Y \sim N(\mu_Y, 1), i = 1, \ldots, n \to \overline{Y}_n$$
$$Z \sim N(\mu_Z, 1), i = 1, \ldots, n \to \overline{Z}_n$$

Admissibility We say an estimator $\hat{\theta}_n$ is admissible if there is no other estimator $\tilde{\theta}$ such that (note that MSE is always a function of a parameter)

$$MSE(\tilde{\theta}) \le MSE(\hat{\theta})$$

i.e. your estimator is always better than someone else's estimator. Now the paradox here is that each pairwise pairs of these 3 (vectors with 2 components, i.e. $(\overline{X}_n, \overline{Y}_n)$) random variables have admissibility, but as soon as you introduce all 3 random variables, you no longer have admissibility.

So we have shown that the MSE of $\overline{X}_n$ is unbiased with minimal assumptions. But we talked

about several estimators. Are there any other estimators other than $\overline{X}_n$? Actually, we have uncountably infinite estimators. Suppose X_i have the same mean μ.

$$\tilde{X}_n = \sum_{i=1}^{n} c_i X_i \text{ where } \sum_{i=1}^{n} c_i = 1$$

$$E(\tilde{X}_n) = E\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} E[c_i X_i] = \sum_{i=1}^{n} c_i E(X_i) = \sum_{i=1}^{n} c_i \mu = \mu \sum_{i=1}^{n} c_i = \mu$$

$$MSE(\tilde{X}_n) = Var(\tilde{X}_n) + \underbrace{Bias^2(\tilde{X}_n)}_{0}$$

Next class we will show that we can minimize the variance if we set $c_i = \frac{1}{n}$
Min $MSE(\tilde{X}_n)$ with $c = (c_1, \ldots, c_n)$ such that $\sum_{i=1}^{n} c_i = 1$

4 01/18/18

Recall We we're talking about number of estimators and showed that there can be uncountably infinite estimators. So the question was, how can we chose amongst uncountably infinite estimators? So we assumed that we wanted to confine ourselves to unbiased estimators. We then looked at sample average, which gives equal weight to each sample. A familiar example is GPA, which has weight based on the amount of credits the course is. We normalize the sample average by making the sum add up to 1. We noticed that if all our observations have the same average μ, then so does our estimator. So how do we chose our estimator?

$$\tilde{X}_{n,\vec{c}} = \sum_{i=1}^{n} c_i x_i \text{ where } \vec{c} = (c_1, \ldots, c_n)$$

$$\{\tilde{X}_{n,\vec{c}} : \vec{c} \in \mathbb{R}^n, \sum_{i=1}^{n} c_i = 1\}$$

One of the things we used to check how good an estimator was is MSE:

$$MSE(\vec{X}_{n,\vec{c}}) = Var(\tilde{X}_{n,\vec{c}}) + \underbrace{[Bias(\tilde{X}_{n,\vec{c}})]^2}_{E(\tilde{X}_{n,\vec{c}}) - \mu}$$

$$MSE(\tilde{X}_{n,\vec{c}}) = Var(\tilde{X}_{n,\vec{c}})$$

So we want:

$$\min_{\vec{c}\in\mathbb{R}} Var(\tilde{X}_{n,\vec{c}}) \tag{1}$$

subject to $\sum_{i=1}^{n} c_i = 1$ (constraint)
This is a generalization of problem 8.6 on page 394.
Suppose X_i's have the same mean μ and variance σ^2 and $Cov(X_i, X_j) = 0, i \neq j$ (i.e. they are orthogonal). Then the solution to (1) is $\overline{X}_n$, i.e. $c_i = \frac{1}{n}, i = 1, \ldots, n$. Now how do we solve this? We can use calculus to find a minimum. How many variables do we have here? $n-1$ variables, as the last one is specified by the constraint. The way we do this is with the Lagrange Method.

So we note that problem (1) is equivalent to

$$\min_{\vec{c}\in\mathbb{R}} \{Var(\tilde{X}_{n,\vec{c}}) + \lambda \left(\sum_{i=1}^{n} c_i - 1 \right)\} \tag{2}$$

where λ is the lagrange multiplier. How do we show that these two mathematical programs are the same? We show that if we find a $\vec{c^*}$ that minimizes (1), it must also minimize (2) and vice versa. How do we solve this?

$$\begin{aligned} Var(\tilde{X}_{n,\vec{c}}) = Var\left(\sum_{i=1}^{n} c_i X_i \right) &\overset{\text{Thm 5.12(b), p271}}{=} \sum_{i=1}^{n} Var(c_i X_i) + 2\sum\sum_{1\leq i\leq j\leq n} Cov(c_i X_i, c_j X_j) \\ &= \sum_{i=1}^{n} c_i^2 Var(X_i) + 2\sum\sum_{1\leq i\leq j\leq n} c_i c_j Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} c_i^2 \overbrace{Var(X_i)}^{\sigma^2} = \sigma^2 \sum_{i=1}^{n} c_i^2 \end{aligned}$$

Now our problem (2) is equivalent to

$$\min_{\vec{c}\in\mathbb{R}} \underbrace{\left\{ \sigma^2 \sum_{i=1}^{n} c_i^2 + \lambda \left(\sum_{i=1}^{n} c_i - 1 \right) \right\}}_{\varphi(\vec{c})}$$

$$\begin{aligned} \frac{\partial}{\partial c_i} \varphi(\vec{c}) &= 2\sigma^2 c_i + \lambda, i = 1, 2, \ldots, n \\ \frac{\partial}{\partial \lambda} \varphi(\vec{c}) &= \sum_{i=1}^{n} c_i - 1 \end{aligned}$$

So now we equate them to 0:

$$\begin{cases} \frac{\partial}{\partial c_i}\varphi_\lambda(\vec{c}) = 0, i = 1, \dots, n \\ \frac{\partial}{\partial \lambda}\varphi_\lambda(\vec{c}) = 0 \to \sum_{i=1}^n c_i = 1 \end{cases}$$

$$\begin{aligned} \frac{\partial}{\partial c_i}\varphi_\lambda(\vec{c}) &= 2\sigma^2 c_i + \lambda = 0 \\ c_i &= -\frac{\lambda}{2\sigma^2}, i = 1, 2, \dots, n \\ \frac{\partial \varphi_\lambda(\vec{c})}{\partial \lambda} &= \sum_{i=1}^n c_i - 1 = 0 \implies \sum_{i=1}^n -\frac{\lambda}{2\sigma^2} = 1 \\ \lambda &= \frac{1}{-\sum_{i=1}^n \frac{1}{2\sigma^2}} = -\frac{2\sigma^2}{n} \end{aligned}$$

$$\begin{cases} c_i = -\frac{\lambda}{2\sigma^2}, i = 1, \dots, n \\ \lambda = -\frac{2\sigma^2}{n} \end{cases}$$

$$\begin{aligned} \to c_i &= \frac{1}{n}, i = 1, \dots, n \\ \vec{c} &= (c_1, c_2, \dots, c_n) \end{aligned}$$

How do we know this is the minimum? We could take the second derivative and look at the resulting matrix, but we won't be going into details for that in this class.

You don't need to check the matrix, but you should be able to take partial derivatives to minimize something (might be less variables, like 8.6), add the constraint to the objective function.

Essentially what we did was minimize $MSE(\tilde{X}_{n,\vec{c}})$. For any vector $\vec{c}$ we have an estimator and this estimator has a distance, so we want to minimize the distance by choosing the components of $\vec{c}$ subject to $\sum_{i=1}^n c_i = 1$ and we found that the solution is the sample average where we give equal weight to each sample, i.e.

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$$

with $c_i = \frac{1}{n}, i = 1, \dots, n$.

So we have constrained ourselves to unbiased estimators that are linear combinations: $\tilde{X}_{n,\vec{c}} = \sum_{i=1}^n c_i X_i$

How do we know that there isn't a better class of distributions that aren't linear combinations? We will see something in chapter 9.

So far we have looked at $E(X) = \mu$ for population averages. But we can also look at other parameters that are important to us, like variance.
So we have a bunch of estimates and want to estimate the variation. The estimation of this variation is very important. $Var(X) = \sigma_x^2$, because sample size requires this and greatly affects the amount of money required to conduct an experiment. How can we solve this? Suppose we have a sample from a population (no distribution assumption):

$X_1, \ldots, X_n$
$E(X_i) = \mu, i = 1, \ldots, n$
$Var(X_i) = \sigma^2, i = 1, \ldots, n$
$Cov(X_i, X_j) = 0, i \neq j$

What would be a natural estimator for this variance? **Sample variance.**

$$\begin{aligned} Var(X) &= E[(X-\mu)^2] \\ S_{n,*}^2 \text{ (sample variance)} &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \end{aligned}$$

Do we have the following equality? This would imply unbiasness.

$$\begin{aligned} E[S_{n,*}^2] &\stackrel{?}{=} \sigma^2 \\ (X_i-\mu)^2 &= [(X_i - \overline{X}_n) + (\overline{X}_n - \mu)]^2 \\ (X_i-\mu)^2 &= (X_i - \overline{X}_n)^2 + (\overline{X}_n - \mu)^2 + 2(X_i - \overline{X}_n)(\overline{X}_n - \mu), i = 1, \ldots, n \\ \sum_{i=1}^{n}(X_i,\mu)^2 &= \sum_{i=1}^{n}(X_i - \overline{X}_n)^2 + \sum_{i=1}^{n}(\overline{X}_n - \mu)^2 + 2\sum_{i=1}^{n}(X_i - \overline{X})(\overline{X}_n - \mu) \\ &= \sum_{i=1}^{n}(X_i - \overline{X}_n)^2 + n(\overline{X}_n - \mu)^2 + 2(\overline{X}_n - \mu)\underbrace{\sum_{i=1}^{n}(X_i - \overline{X}_n)}_{\sum_x i - n\overline{X}_n = 0} \\ \sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}(X_i - \overline{X}_n)^2 + n(\overline{X}_n - \mu)^2 \\ E\left[\sum_{i=1}^{n}(X_i-\mu)^2\right] &= E\left[\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right] + E[n(\overline{X}_n - \mu)^2] \end{aligned}$$

$$\sum_{i=1}^{n} E(X_i - \mu)^2 = E[nS^2 n, *] + E[n(\overline{X}_n - \mu)^2]$$

$$\sum_{i=1}^{n} \underbrace{Var(X_i)}_{\sigma^2} = nE(S^2_{n,*}) + nE[(\overline{X}_n - \mu)^2]$$

$$n\sigma^2 = nE(S^2_{n,*}) + n\underbrace{E[(\overline{X}_n - \mu)^2]}_{Var(\overline{X}_n)}$$

$$\sigma^2 = E(S^2_{n,*}) + Var(\overline{X}_n)$$

$$\sigma^2 = E(S^2_{n,*}) + \frac{\sigma^2}{n}$$

$$E(S^2_{n,*}) = \sigma^2\left(1 - \frac{1}{n}\right)$$

So is this unbiased? No! It's not equal to our target. When n gets really big, the other term will be negligible, but what if we want something unbiased without that condition? Multiply by the reciprocal!

$$\frac{n}{n-1}E(S^2_{n,*}) = \sigma^2$$

$$E\left(\frac{n}{n-1}S^2_{n,*}\right) = \sigma^2$$

$$S^2_n = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

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

So now we have something unbiased. But why is sample variance happen to be biased? Why do we need $n-1$? What is causing us trouble? $\overline{X}_n$. If we had μ instead, we wouldn't have this problem.

Consider the following as vectors:

$$V = Span\{X_i - \overline{X}_n, i = 1, \ldots, n\}, \sum_{i=1}^{n}(X_i - \overline{X}_n) = 0$$

Not all the vectors are linearly independent, so we lose 1 degree of freedom, the reason why we have $n-1$. But now, if we compare this with:

$$W = Span\{X_i - \mu, i = 1, \ldots, n\}$$

$$dim(V) = n - 1, dim(W) = n$$

The X_i do not depend linearly on μ, if you sum up all the terms it won't be 0, because μ is the population average, not the sample average, we are just shifting by a constant μ and this will not affect covariance.

So far we have looked at the parameters: μ, p, σ^2.
What happens if we have 2 populations, i.e. men $(X_1, \ldots, X_n)$ with mean μ_M and women $(Y_1, \ldots, Y_n)$ with mean μ_W. Say we want to compare the salaries of men and women. We may have the salaries of everyone in our university, but not everyone in Canada. So what's a natural estimator for μ_M and μ_w?

$$\overline{X}_M = \frac{1}{m}\sum_{i=1}^{m} X_i$$
$$\overline{Y}_W = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Exercise Show that $E(\overline{X}_M - \overline{Y}_W) = \mu_M - \mu_W$ (this should be immediately available to you from what we've done). Note that we assume they are all coming from the same population (i.e. they all have the same mean).
Furthermore, also show:
Suppose $Cov(X_i, X_j) = 0, Cov(Y_i, Y_j) = 0, i \neq j$ and Xs and Ys are independent (don't really need this condition, extra). Find $Var(\overline{X}_m - \overline{Y}_W)$ (these are all immediately available to you, just work through them)
The formula that will be useful to you (Thm 5.12):

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) \pm 2abCov(X, Y)$$

The same idea applies if we want to see the proportion of a kind of population that votes for someone and more applications.