# MATH 324: Statistics

Julian Lore

Last updated: February 6, 2018

Notes from Masoud Asgharian's Winter 2018 lectures.

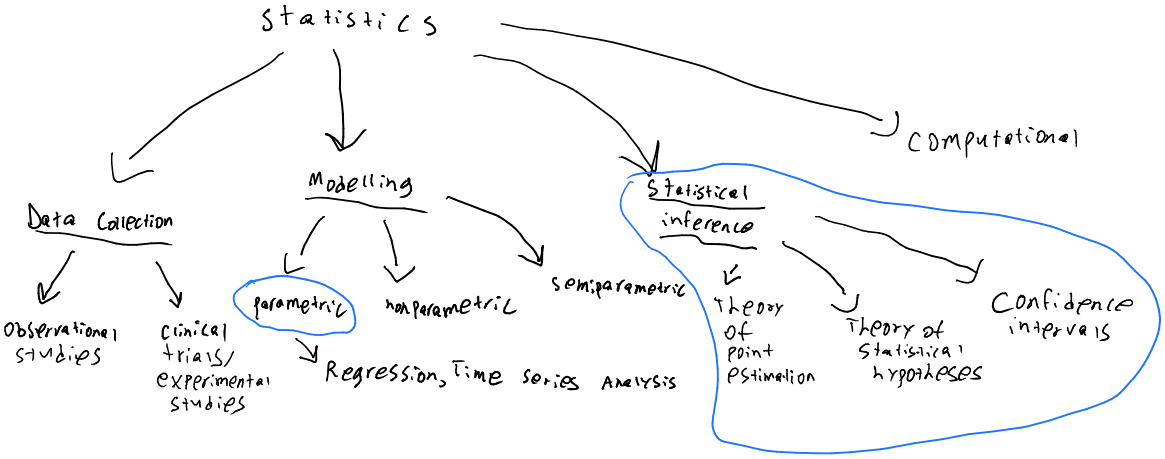
## Contents

# 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 $\mu$ 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 $\mu$. 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 $\mu$, 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 $\mu$, 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 $\mu$ moves from $-\infty$ to $\infty$, 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 $\mu$ 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 $\varepsilon$ 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 $\lambda$ 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 $X_i \sim \text{Bernouilli}(p)$, 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 $\varepsilon, \delta$ 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:

$$\perp\!\!\!\perp_{i=1}^{n} X_i$$
$$\hat{p}_n = \frac{1}{n} \underbrace{\sum_{i=1}^{n} x_i}_{\substack{\text{Number of people} \\ \text{in sample voting NDP}}} = \overline{X}_n$$
$$\varepsilon = |\hat{p}_n - p|$$
$$p(|\hat{p}_n - p| > \overbrace{\delta}^{\text{given}}) \leq \frac{E[|\hat{p}_n - p]^2}{\delta^2}$$

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

$$
\begin{aligned}
E(X_i) &= 1 \cdot p(X_i = 1) + 0 \cdot p(X_i = 0) \\
&= 1 \cdot p + 0 \cdot (1 - p) \\
&= p(\text{Expected value of a Bernoulli variable}) \\
\implies E(\hat{p}_n) &= \frac{1}{n}\sum_{i=1}^{n} p = \frac{1}{n}np = p
\end{aligned}
$$

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:

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

$$= 1 \cdot p + 0 \cdot (1-p) = p$$

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

Now, using Chebyshev's inequality:

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

Is this useful? As $n$ tends to $\infty$, the bound goes to 0, meaning, no matter what $\delta$ 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.

$$\begin{aligned} f\left(\frac{1}{2}\right) &= \frac{1}{2}\left(1 - \frac{1}{2}\right) = \frac{1}{4} \\ p(|\hat{p}_n - p| > \delta) &\leq \frac{1}{4n\delta^2} \end{aligned}$$

So now for any $\delta$ 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.

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

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 $\mu$. 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 $\theta$ 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 $\theta$ if $E(\hat{\theta}_n) = \theta$.

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

$$\begin{aligned}
\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i &\implies E(\overline{X}_n) = \mu \\
Var(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n} Var(X_i) &= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \\
MSE(\overline{X}_n) = Var(\overline{X}_n) + [Bias(\overline{X}_n)]^2 &= \frac{\sigma^2}{n} + 0^2 = \frac{\sigma^2}{n}
\end{aligned}$$

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 $\mu$. 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

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

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

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

where $\sigma^2 = Var(X_i), i = 1, \ldots, n$. i.e. if $X_i's$ have the same mean $\mu$ & Variance $\sigma^2$ 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)

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

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 $\mu$.

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

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

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 $\mu$, 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 $\mu$ and variance $\sigma^2$ 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 $\lambda$ 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?

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

$$= \sum_{i=1}^{n} c_i^2 Var(X_i) + 2\sum\sum_{1\leq i\leq j\leq n} c_i c_j Cov(X_i, X_j)$$

$$= \sum_{i=1}^{n} c_i^2 \overbrace{Var(X_i)}^{\sigma^2} = \sigma^2 \sum_{i=1}^{n} c_i^2$$

Now our problem (2) is equivalent to

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

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

So now we equate them to 0:

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

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

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

$$\to c_i = \frac{1}{n}, i = 1, \ldots, n$$
$$\vec{c} = (c_1, c_2, \ldots, c_n)$$

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 $c_i = \frac{1}{n}, i = 1, \ldots, n$.
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.**

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

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

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

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!

$$
\begin{aligned}
\frac{n}{n-1}E(S_{n,*}^2) &= \sigma^2 \\
E\left(\frac{n}{n-1}S_{n,*}^2\right) &= \sigma^2
\end{aligned}
$$

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

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 $\mu$ 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 $\mu$, if you sum up all the terms it won't be 0, because $\mu$ is the population average, not the sample average, we are just shifting by a constant $\mu$ and this will not affect covariance.

---

So far we have looked at the parameters: $\mu, p, \sigma^2$.
What happens if we have 2 populations, i.e. men $(X_1, \ldots, X_n)$ with mean $\mu_M$ and women $(Y_1, \ldots, Y_n)$ with mean $\mu_W$. 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 $\mu_M$ and $\mu_w$?

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

**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 $X$s and $Y$s 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.

# 5 01/15/18

## 5.1 Confidence Intervals

Statistics is inductive logic, meaning that we have a sample of the whole population. Using this sample we want to make inference about the whole population. This is unlike theorems that are deductive, where we go to specifics from general things, we are going up, generalizing the sample to the whole population.

Suppose we have $X_i \overset{iid}{\sim} N(\mu, 1), i = 1, \ldots, n$ (IQ of high school graduates). We want to estimate the IQ for the whole population of Canada.

We've learned that $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a reasonable point estimate. But we want to do an interval estimate.

**Pivotal quantity** A function of the observations $X_1, \ldots, X_n$ and some unknown parameters, ideally just the parameters of interest, such that the distribution of this function does **NOT** depend on any unknown parameter.

What do we mean? What is the distribution of $\overline{X}_n$?
$\overline{X}_n \sim N(\mu, \frac{1}{n})$. How do we get this? Remember that we had different approaches to getting distributions of a function of random variables.

- Transformation (essentially the change of variables theorem from Calculus, when you integrate a function and change the variable).
- Method of distribution (try to relate the distribution to an existing distribution)

- Method of mgf (most suitable when we have iid samples and want to find the distribution of the sum of random variables, exactly what we have here with $\overline{X}_n$)

How does the method of mgf work?

$$S = \sum_{i=1}^{n} X_i$$

$$m_S(t) = E[e^{tS}] = E\left[e^{t\sum_{i=1}^{n} X_i}\right] = E\left[\prod_{i=1}^{n} e^{tX_i}\right]$$

$$\text{Independence} \implies = \prod_{i=1}^{n} E[e^{tX_i}] = \prod_{i=1}^{n} m_{X_i}(t)$$

$$\text{Identically dist} \implies = \prod_{i=1}^{n} m(t) = [m(t)]^n$$

For example, if $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ then

$$m_{X_i}(t) = \left[e^{\mu t + \frac{\sigma^2 t^2}{2}}\right]^n = e^{n\mu t + \frac{n\sigma^2 t^2}{2}}$$

Now if we call $n\mu = \mu_*$ and $n\sigma^2 = \sigma_*^2$

$$= e^{\mu_* t + \frac{\sigma_*^2 t^2}{2}}$$

Now this is the same mgf as a normal distribution with

$$S \sim N(\mu_*, \sigma_*^2) = N(n\mu, n\sigma^2)$$

$$\overline{X}_n = \frac{1}{n} S$$

$$m_{\overline{X}_n}(t) = E[e^{t\overline{X}_n}] = E[e^{t\frac{1}{n}S}] \overset{t\frac{1}{n}=t_*}{=} E[e^{t_* S}] = m_S(t_*) = e^{n\mu t_* + \frac{n\sigma^2 t_*^2}{2}}$$

$$= e^{n\mu\frac{t}{n} + \frac{n\sigma^2\left(\frac{t}{n}\right)^2}{2}} = e^{\mu t + \frac{\sigma^2 t^2}{2n}}$$

Now this tells us $\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ if $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$.

In our example, $\sigma^2 = 1$, so $\overline{X}_n \sim N(\mu, \frac{1}{n})$. Standardizing we get:

$$Z = \underbrace{\frac{\overline{X}_n - \mu}{\sqrt{\frac{1}{n}}}}_{\text{Pivotal quantity}} \sim N(0,1)$$

This distribution itself does not depend on any unknown parameter, it is just the standard normal distribution. All standardizations of normal distributions make them into pivotal quantities.

So we know that:

$$P(|Z| \leq 1.96) = 0.95$$

Now let's try and plug in for $Z$:

$$P\left(\left|\frac{\overline{X}_n - \mu}{\sqrt{\frac{1}{n}}}\right| \leq 1.96\right) = 0.95$$

$$P\left(\overline{X}_n - 1.96\sqrt{\frac{1}{n}} \leq \mu \leq \overline{X}_n + 1.96\sqrt{\frac{1}{n}}\right) = 0.95$$

This gives me an interval estimate that $\mu$ falls between these two limits with 95% confidence. Why were we able to do this? Because $Z$ is a pivotal quantity so we could just use the table. If $Z$ depended on unknown parameters we would need to find them. So we have found an upper and lower bound with a given confidence.

Now, we can generalize this. $\left(\overline{X}_n \pm 1.96\sqrt{\frac{1}{n}}\right)$ is called a 95% confidence interval, for the unknown $\mu$. In real life application we have one sample and one average. So we plug in values and get the two bounds, say: $(125, 135)$. Someone might come in and ask us how we can associate probability to this interval, either the average is in texthere or not, so what do we mean by 95% confidence? Notice that the probability statement we had above means that, if we dispatch many students to test students and come up with confidence intervals just like we did and write them all on pieces of paper. Then we randomly select one of the papers from the bag, then we are 95% sure that the interval given **covers the truth**.

Let's see how we can generalize this. In general, we have large sample confidence intervals. When we started this example, we assumed that the observations are coming from a Normal distribution, but in practice, we don't know if something is normally distributed. This allowed us to normalize $\overline{X}_n$. So if the $X_i$ weren't normally distributed, all of this would fail.

But we know that, from Central Limit Theorem (most celebrated result in all of probability & statistics, in baby form): If $X_i$'s are independent, have the same mean $\mu$ and variance $\sigma^2$, then $\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \overset{approx}{\sim} N(0,1)$ for large $n$. Note, we **do not** need to know the distribution of

$X_i$. This is a pivotal quantity. In many applications, we might only be interested in $\mu$, but we have another unknown parameter, $\sigma$ sitting here. We could do the same calculation as above and get:

$$\left(\overline{X}_n \pm 1.96\sqrt{\frac{\sigma^2}{n}}\right) \text{ is 95\% CI}$$

$$\left(\overline{X}_n \pm 1.96\frac{\sigma}{\sqrt{n}}\right)$$

But how do we get rid of $\sigma$? We can replace it by the sample $\sigma$, i.e.

$$\left(\overline{X}_n \pm 1.96\frac{S}{\sqrt{n}}\right)$$

Note that we are still assuming very large $n$ (how large is very large will not be addressed in this class).

What justifies replacing $\sigma$ with $S$? Since we have a large sample, then $S$ is going to be close enough to $\sigma$. This is called **consistency** (to be discussed in the next chapter). But what justifies replacing a parameter by a consistent estimator?

Essentially, what we are saying is:

$$\frac{\overline{X}_n - \mu}{\frac{S}{\sqrt{n}}} \overset{approx}{\sim} N(0,1)$$

But CLT only works with $\sigma$. Intuitively we say it works when $S$ is sufficiently close to $\sigma$.

$$\frac{\overline{X}_n - \mu}{\frac{S}{\sqrt{n}}} = \frac{\sigma}{S} \cdot \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

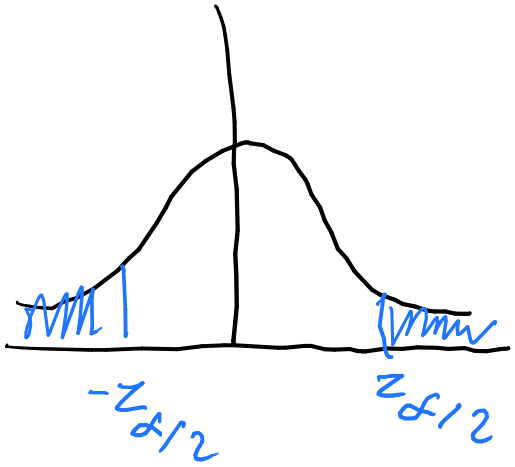
We have a theorem from Cramer, telling us that if we have a ratio of parameters that converge to 1 and a distribution converging to Normal, then the whole thing converges to Normal.

So we can use:

$$\left(\overline{X}_n \pm 1.96\frac{S}{\sqrt{n}}\right) \text{ 95\% CI}$$

So if we need other confidence intervals:

$$\left(\overline{X}_n \pm Z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right) 100(1-\alpha)\% \text{ CI}$$




Note that this is just for $\overline{X}_n$. What if we have something different?

If $n$ is large and $\theta$ is a parameter of interest, then

$$\frac{\hat{\theta}_n - \theta}{\sqrt{Var(\hat{\theta}_n)}} \overset{approx}{\sim} N(0,1)$$

where $\hat{\theta}_n$ is an MLE (Maximum Likelihood Estimator, which we'll see in a future chapter)

For example, if $X_i \sim Ber(p)$, i.e. $X_i = \begin{cases} 1 & p \\ 0 & 1-p \end{cases}$ and $E(X_i) = p$, then $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i = \hat{p}_n$. So we want to estimate $\hat{p}_n$.

$$\frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

$$\hat{p}_n \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

But once again, we don't know what $p$ is and it is not useful! So we can either:

- Replace $p$ by $\hat{p}_n$
- Or, take the conservative approach and replace $p(1-p)$ by $\frac{1}{4}$ (remember that this is the max value the it can take, as seen earlier)
  $f(x) = x(1-x) \to f'(x) = 1-2x = 0 \implies x = \frac{1}{2}, f''(x) = -2, f(\frac{1}{2}) = \frac{1}{4}$, a max. Since we are replacing by the max, we are accounting for the worse situation, so it gives us

a large confidence interval (which is why it is conservative). These are the types of intervals that we hear about before elections. When they are talking about chance they are talking about $\hat{p}_n$, when talking about how accurate it is, they are referring to the $\pm$ part and when saying it is right 19 out of 20 times, they are talking about the % of the CI.

Now suppose we are interested in estimating the variance $\theta = p(1-p)$ (not interested in estimating $p$). What would be a natural estimator?

$$\hat{\theta}_n = \hat{p}_n(1-\hat{p}_n)$$

$$\frac{\hat{\theta}_n - \theta}{\sqrt{Var(\hat{\theta}_n)}} \sim N(0,1)$$

We then have the general recipe:

$$\hat{\theta}_n \pm Z_{\frac{\alpha}{2}}\sqrt{Var(\hat{\theta}_n)} \text{ is } 100(1-\alpha)\% \text{ CI}$$

The only thing we need is a right estimator that gives us this asymptotic normality and then we are in business! As long as our $\hat{\theta}_n$ can provide asymptotic normality for large $n$, then wen can have this. Most of the estimators in the textbook follow this pattern.

---

## 5.2 Small Sample Size

There are some cases we can solve and some that we cannot. The case of interest that we can solve, is the normal case. We are focusing a lot on normal distributions because we will see later that many things relate to the normal distribution, like comparing control groups on placebo and not on placebo. $X_i \overset{iid}{\sim} N(\mu, \sigma^2), i = 1, \ldots, n$ where $n$ is "small". Here $\mu$ is of interest but $\sigma^2$ is a nuisance.

We learned that $\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \overset{\text{Exact}}{\sim} N(0,1)$. So we can use this as a pivotal quantity to come up with a confidence interval for $\mu$. But we still have the $\sigma$ here that is a nuisance. Before we replaced $\sigma$ by $S$, but that was well justified because our sample size was large. But that's the best thing we have, so we'll use it. But can we say it is normally distributed? We were

only able to say that earlier because we had a large distribution. So what might happen?

$$\frac{\overline{X}_n - \mu}{\frac{S}{\sqrt{n}}} \sim T_{n-1}$$

This will be similar to the $X_i$, i.e. normally distributed and it will still have a bell curve. But what about the tails? The uncertainty of $S$ means that the tails will die out more slowly.

If $X_i \overset{iid}{\sim} N(\mu, \sigma^2), i = 1, \ldots, n$, then

$$\frac{\overline{X}_n - \mu}{\frac{S}{\sqrt{n}}} \sim T_{n-1}$$

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

What is the $T$ distribution? $\frac{N(0,1)}{\sqrt{\frac{\chi^2_\nu}{\nu}}}$ where $N(0,1) \perp\!\!\!\perp \chi^2_\nu$ and $\nu$ denotes the degrees of freedom. How does this relate to what we have?

$$\frac{(\overline{X}_n - \mu)}{\frac{\sigma}{\sqrt{n}}} \Big/ \left( \left( \frac{\sigma}{\sqrt{n}} \right)^{-1} \sqrt{\frac{S^2}{n}} \right)$$

Where the numerator is

$$\sim N(0,1)$$

$$= \frac{\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{S^2}{\sigma^2}}}$$

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

$$\sum_{i=1}^n \left( \frac{X_i - \overline{X_n}}{\sigma} \right)^2 \sim \chi^2_{n-1}$$

# 6 01/30/18

**Recall** When it comes to confidence intervals, the main tool is **pivotal quantities**. So whatever we did using large or small samples, was to come up with a pivotal quantity. In large samples, we used Central Limit Theorem. For small samples, the case that we studied assumed that observations came from a Normal distribution and so we used normality

assumption in order to come up with a pivotal quantity. What happens when we go beyond Normal distribution? There is a general approach to get a pivotal quantity that covers all continuous random variables, but in practice it is only useful if we can find the distribution function of the variable of interest, but it is only useful if its easy to acquire. We will present the recipe today, which is sometimes successful and sometimes unsuccessful. Pivotal quantities are case by case, vary by distribution. This is already in the notes on myCourses:

Say we have $X_1, \ldots, X_n \sim F(cdf) \to f(pdf)$. Suppose this is information from Revenue Canada and we made a histogram of the range of salaries and we fit a distribution over several years (this is our pdf). Now for this year, we'll be taking random people ($x_i$) from Canada who have filled out their tax forms.

Probability Integral Transform is a simple result that is very useful, especially in simulation. What it says is that if $X(\text{continuous}) \sim F \implies \underbrace{F(X)}_{Y} \sim Unif(0,1)$. I.e. apply $F$ on a continuous distribution and get another distribution, a uniform distribution on $(0,1)$. So how can we use this? If we can generate observations on a uniform distribution, then we can generate results for any continuous distribution by applying $F^{-1}$, (since $F$ is a monotone decreasing function it has an inverse) and then we get $X$s that are distributed like $X$, although uniform randomness is very difficult to achieve.

$$
\begin{aligned}
&X \sim N(0,1) \\
&\qquad f_X(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}, -\infty < x < +\infty \\
&\qquad F_X(x) = \int_{-\infty}^{x} f_x(t)\, dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt
\end{aligned}
$$

This is a monotone function, but its inverse is very hard to find. What we do is generate uniform random numbers to approximate this and then get normal distribution. Try this at home:

$$X \sim FY = F(x)$$

You will see that the distribution of $Y$ is uniform. Use the method of transformation here (look in Chapter 6). Now our result is:

$$X\ F \implies \underbrace{Y = F(X) \sim Unif(0,1)}_{PIT} \implies -2\log Y \sim \chi^2_2$$

$$X_i \overset{iid}{\sim} F, i = 1, \dots, n$$
$$Y_i = F(X_i) \overset{iid}{\sim} Unif(0,1), i = 1, 2, \dots, n$$
$$V_i = -2 \log Y_i \overset{iid}{\sim} \chi^2_2, i = 1, \dots, n$$

$$\sum_{i=1}^{n} V_i \sim \chi^2_{2n} \tag{3}$$

This can easily be gotten to via the method of mgf. If $F$ has any unknown parameters, it will still be present in $V_i$. So $\sum_{i=1}^{n} V_i$ will depend on $X_i$s and parameters of $F$, but its distribution doesn't, so this is a pivotal quantity.

**Ex** $X_i \overset{iid}{\sim} Exp(\lambda), i = 1, 2, \dots, n$

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases} \lambda > 0$$
$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt = \int_0^x \lambda e^{-\lambda t}\, dt$$
$$= -e^{-\lambda t}|_0^x = 1 - e^{-\lambda x}$$

The dual to (3):

$$\sum_{i=1}^{n} W_i \sim \chi^2_{2n}$$

where $W_i = -2 \log(1 - Y_i)$. This is because, if $U \sim Unif(0,1)$ then $1 - U \sim Unif(0,1)$. This is useful because we often want $1 - F$, because the probability of $F$ is that something is less than something, so $1 - F$ means that the value is past a certain point, i.e. this is good for the survivor function, checking the chances that someone survives past something.

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 - e^{-\lambda x} & \text{if } x > 0 \end{cases}$$
$$S_X(x) = 1 - F_X(x) = \begin{cases} 1 & \text{if } x \leq 0 \\ e^{-\lambda x} & \text{if } x > 0 \end{cases}$$

$$\begin{aligned}\sum_{i=1}^{n} W_i &= \sum_{i=1}^{n} -2\log[1-F(X_i)] \\ &= \sum_{i=1}^{n} -2\log e^{-\lambda X_i} \\ &= 2\lambda \sum_{i=1}^{n} X_i = 2\lambda n\overline{X}_n\end{aligned}$$

$$2\lambda n\overline{X}_n = \sum_{i=1}^{n} W_i \sim \chi^2_{2n}$$

This is a pivotal quantity. So we can use this for a confidence interval of lambda.

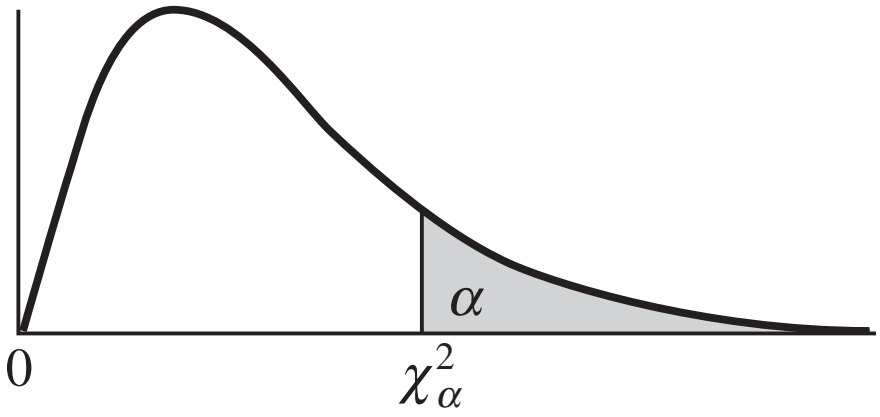
Using the $\chi^2$ table (Appendix 3, p850-851), we can find:

$$\chi^2_{2n,0.025} \& \chi^2_{2n,0.975}$$

$$P\left(\chi^2_{2n,0.975} < 2\lambda n\overline{X}_n < \chi^2_{2n,0.025}\right) = 95\% = 0.95$$

$$P\left(\frac{\chi^2_{2n,0.975}}{2n\overline{X}_n} < \lambda < \frac{\chi^2_{2n,0.025}}{2n\overline{X}_n}\right) = 95\% = 0.95$$

So: $\left(\frac{\chi^2_{2n,0.975}}{2n\overline{X}_n}, \frac{\chi 2n,0.025}{2n\overline{X}_n}\right)$ is a 95% C.I. for $\lambda$.




## 6.1 Sample Size Determination

Say we want to predict the results of an election. We want to be, say 99% sure of the results (why not 100%? because then we'd need infinite samples).

$$X_i = \begin{cases} 1 & \text{NDP} \\ 0 & \text{otherwise} \end{cases}, i = 1, \ldots, n$$

$$p(X_i = 1) = p$$

$$X_i \overset{iid}{\sim} Bernoulli(p), i = 1, \ldots, n$$

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

We want $|\hat{p}_n - p| \leq 0.01$ with 95%. We can use Chebyshev's for this, but since Chebyshev's makes very little assumptions, the bound is very crude. We want a tighter bound. So we can use CLT:

$$\hat{p}_n \pm \underbrace{1.96\sqrt{\frac{p(1-p)}{n}}}_{B}$$

This is called a symmetric confidence interval, since it is centered around a point. Normal distribution is symmetric, $\chi$ is not. If it is symmetric, we can talk about length or half length, if it isn't we talk about the whole length.

$$(\text{e.g. } 0.01) \to B = 1.96\sqrt{\frac{p(1-p)}{n}}$$

Say we want it in terms of $100(1-\alpha)\%$:

$$B = Z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$$

We obtain:

$$n = \frac{p(1-p)(Z_{\frac{\alpha}{2}})^2}{B^2}$$

We have the following 2 options:

- Replace $p$ by $\hat{p}_n$ (a prior estimate, although often we don't have an estimate).
- Replace $p(1-p)$ by $\frac{1}{4}$ (conservative approach).

So then we get:

$$n = \frac{Z^2_{\frac{\alpha}{2}}}{4B^2}$$

(we round up $n$ if it is not an integer) This gives us a sample size! This is a proportion estimate for the sample size for Bernoulli.

**Ex** $B = 0.01, \alpha = 0.05$

$$n = \frac{(1.96)^2}{4(0.01)^2} = 9604 \approx 10,000$$

---

Now, generalizing:

C.I. for $\mu$:

$$\overline{X}_n \pm \underbrace{Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}}_{B}$$

$$B = Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$n = \frac{Z_{\frac{\alpha}{2}}^2 \sigma^2}{B^2} \text{ for } \mu$$

Here we can't just maximize $\sigma$, so we do pilot studies, where we try to justify the need for a study, so we receive a little bit of money to make a small study, to get an idea of some of the parameters, like $\sigma$ in this case and then we use this parameter to get a bit of information on other parameters so that we can apply for the real study. Now determining the sample size goes back to confidence intervals.

---

Other variables:

$$\frac{\hat{\theta}_n - \theta}{\sqrt{Var(\hat{\theta}_n)}} \overset{app}{\sim} N(0,1) \text{ for large } n \tag{4}$$

Here $\hat{\theta}_n$ is the MLE (will see in the future) and $\theta$ is the estimand.

$$\hat{\theta}_n \pm Z_{\frac{\alpha}{2}} \sqrt{Var(\hat{\theta}_n)}$$

$100(1-\alpha)\%$ C.I. for $\theta$

$$B = Z_{\frac{\alpha}{2}} \sqrt{Var(\hat{\theta}_n)}$$

So we get:

- Bernoulli $Var(\hat{\theta}_n) = \frac{p(1-p)}{n}$
- $\mu$ $Var(\overline{X}_n) = \frac{\sigma^2}{n}$

These are based on a large sample. What if we have a small sample size? As long as we have the confidence interval, if it is symmetric we can bound by half length, if it isn't we can bound by the whole length.

Normal case:

$$\overline{X}_n \pm t_{\frac{\alpha}{2},(n-1)} \frac{\sigma}{\sqrt{}}$$
$$B = t_{\frac{\alpha}{2},(n-1)} \frac{s}{\sqrt{n}}$$

---

So far we've talked about one population. So what happens if we have two populations? Say we have men and women and want to estimate average salary.

$$\frac{\hat{\theta}_n - \theta}{\sqrt{Var(\hat{\theta}_n)}} \sim N(0,1)$$

How do we use this if we have more than one sample, say $X_1, \ldots, X_m$ (men, $\overline{X}_m$) and $Y_1, \ldots, Y_n$ (women, $\overline{Y}_n$). Say we want to estimate $\theta = \mu_M - \mu_W$. Then $\hat{\theta} = \overline{X}_m - \overline{Y}_n$. Now, using the formula above, we get:

$$\frac{(\overline{X}_m - \overline{Y}_m) - (\mu_M - \mu_W)}{\sqrt{Var(\overline{X}_m - \overline{Y}_m)}}$$

So the general recipe is:

Pivotal quantity -¿ confidence interval -¿ margin of error -¿ sample size

$$Var(\overline{X}_m - \overline{Y}_n) = Var(\overline{X}_m) + Var(\overline{Y}_n) - 2Cov(\overline{X}_m, \overline{Y}_n)$$

Assuming that $X$s & $Y$s are independent (note that we don't need independence, just orthogonality such that $Cov$ is 0, independence is much stronger). Then $Cov(\overline{X}_m, \overline{Y}_n) = 0$

$$Var(\overline{X}_m - \overline{Y}_n) = Var(\overline{X}_n) + Var(\overline{Y}_n) = \frac{\sigma_M^2}{m} + \frac{\sigma_W^2}{n}$$

$$\frac{(\overline{X}_m - \overline{Y}_n) - (\mu_M - \mu_W)}{\sqrt{\frac{\sigma_M^2}{m} + \frac{\sigma_W^2}{n}}} \overset{approx}{\sim} N(0,1) \text{ for large } m\&n$$

$$(\overline{X}_m - \overline{Y}_n) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_M^2}{m} + \frac{\sigma_W^2}{n}}$$

is a $100(1-\alpha)\%$ C.I. for $\mu_M - \mu_w$

$$B = Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_M^2}{m} + \frac{\sigma_W^2}{n}}$$

$$\left(\frac{B}{Z_{\frac{\alpha}{2}}}\right)^2 = \frac{\sigma_M^2}{m} + \frac{\sigma_W^2}{n}$$

So we have one equation with 2 unknowns. What should we do? Set them equal! What is the rational for setting them equal? Cost. If $u = km$, we can solve:

$$\left(\frac{B}{Z_{\frac{\alpha}{2}}}\right)^2 = \frac{1}{m}\left(\sigma_M^2 + \frac{\sigma_W^2}{k}\right)$$

$$m = \frac{\sigma_M^2 + \frac{\sigma_W^2}{k}}{\left(\frac{B}{Z_{\frac{\alpha}{2}}}\right)^2}$$

With proportions $\sigma_M^2 = p_M(1-p_M)$ can replace with $\frac{1}{u}$, much simpler:

$$\to \frac{1 + \frac{1}{k}}{\left(\frac{B}{Z_{\frac{\alpha}{2}}}\right)^2}$$

# 7 02/01/18

Today we are starting **Chapter 9**.

## 7.1 Relative Efficiency

We learned that we can have many point estimates and we discussed estimator errors with Chebyshev's inequality as well as MSE, etc. We noticed that MSE could be split into 2 parts, bias and variance. So perhaps a reasonable property for an estimator would be that we want it to be unbiased. But we also saw that unbiased estimators are not unique, so which do we chose? We came to the point with two unbiased estimators and decided that

the one with the **smaller variance** is better. Now we want to quantify this deviation by looking at the difference in variances of unbiased estimators, or the ratio. For us, ratios and differences are the same (if you take logarithm of a ratio you get differences). We do prefer to work with ratios though.

**Def** The *relative efficiency* of two **unbiased estimators**, $\hat{\theta}_1$ and $\hat{\theta}_2$, is defined to be

$$eff(\hat{\theta}_1, \hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta})_1}$$

Note that we can also extend this to bias estimators, looking at MSEs instead. But there is a reason that we are using unbiased estimators. For a fixed sample size $n$, we may have biased estimators, but for large $n$, as $n$ increases, we require unbiasedness.

The reason we look at the ratio:

Suppose we have an unfair coin, with $P(head) = 0.6, P(tail) = 0.4$. If you toss this over and over again, you'll see that the ratio of heads and tails should be bounded, converge to some number. But for difference, it won't converge, because the difference increases more and more to infinity.

Recall determining sample size. We looked at the margin of error, the part we'd get in the confidence interval. If we want to compare two different procedures/estimators.

Recall that we had $Var(\overline{X}_n) = \frac{\sigma^2}{n}$. If we have a symmetric distribution we can replace $\overline{X_n}$ by $Var(median_n) = \frac{\sigma^2}{\ldots 1.6n}$. So using a procedure that uses the median requires say, 1.6 times the sample size to get the same information, much more costly. So the ratio tells us about the ratio between two estimators and which one requires a larger sample size for the same information.

**Ex. 9.1, p446** $Y_i \overset{iid}{\sim} Unif(0, \theta), i = 1, 2, \ldots, n$, where $\theta$ is what we want to estimate. Now consider the following 2 estimators:

$$\hat{\theta}_1 = 2\overline{Y}_n$$
$$\hat{\theta}_2 = \left(\frac{n+1}{n}\right) Y_{(n)}$$

where $Y_{(n)} = \max\{Y_1, \ldots, Y_n\}$. Which one intuitively seems better? Try to look at the extreme cases, like when $n = 1$ or $n$ is very large. When $n$ is very large, $\hat{\theta}_2$ gets larger and larger and since $\theta$ is the upper bound, it will get closer and closer to $\theta$, but it still remains

slightly under, which is why we multiply by $\frac{n+1}{n}$ to increase it slightly.

$$\begin{aligned}\hat{\theta}_1 : E(\hat{\theta}_1) &= E(2\overline{Y}_n) = 2E(\overline{Y}_n) = 2E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\ &= \frac{2}{n}E\left(\sum_{i=1}^{n} Y_i\right) = \frac{2}{n}\sum_{i=1}^{n} E(Y_i)\end{aligned}$$

Because $Y$ is a symmetric distribution, the mean is the median:

$$\begin{aligned}&= \frac{2}{n}\cdot\sum_{i=1}^{n}\frac{\theta}{2} = \theta \\ Var(\hat{\theta}_1) &= Var(2\overline{Y}_n) = \frac{4Var(Y)}{n} \\ Var(Y) &= E(Y^2) - \underbrace{[E(Y)]^2}_{\frac{\alpha}{2}} \\ E(Y^2) &= \int_{-\infty}^{\infty} y^2 f_Y(y)\, dy = \int_0^{\theta} y^2\frac{1}{\theta}\, dy = \frac{1}{\theta}\int_0^{\theta} y^2\, dy \\ &= \frac{1}{\theta}\cdot\frac{1}{y}y^3\Big|_0^{\theta} = \frac{\theta^3}{3\theta} = \frac{\theta^2}{3} \\ Var(Y) &= \frac{\theta^2}{3} - \frac{\theta^2}{4} = \frac{\theta^2}{12}\end{aligned}$$

Plugging in:

$$Var(\hat{\theta}_1) = \frac{4\cdot\frac{\theta^2}{12}}{n} = \frac{\theta^2}{3n}$$

$$\begin{aligned}\hat{\theta}_2 : F_{Y_{(n)}} &= P(Y_{(n)} \le t) = P(Y_1 \le t, Y_2 \le 2\ldots., Y_n \le t) \\ &= \prod_{i=1}^{n} P(Y_i \le t) = \prod_{i=1}^{n} F_{Y_i}(t) = [F_Y(t)]^n \\ F_{Y_{(n)}}(t) &= [F_Y(t)]^n \\ \delta_{Y_{(n)}}(t) &= \frac{d}{dt}F_{Y_{(n)}}(t) = \frac{d}{dt}[F_Y(t)]^n \\ &= nf_Y(t)(F_Y(t))^{n-1}\end{aligned}$$

$$
\begin{aligned}
f_{Y_{(n)}}(t) &= \begin{cases} n\frac{1}{\theta}\left(\frac{t}{\theta}\right)^{n-1} & \text{if } 0 < t < \theta \\ 0 & \text{otherwise} \end{cases} \\
E(\hat{\theta}_2) &= E\left[\left(\frac{n+1}{n}\right) Y_{(n)}\right] = \left(\frac{n+1}{n}\right) E[Y_{(n)}] \\
&= \left(\frac{n+1}{n}\right) \int_0^\theta y \underbrace{n \cdot \frac{1}{\theta}\left(\frac{y}{\theta}\right)^{n-1}}_{f_{Y_{(n)}}(y)}\, dy \\
&= \left(\frac{n+1}{n}\right) \cdot \frac{n}{\theta^n} \int_0^\theta y^n \, dy \\
&= \left(\frac{n+1}{n}\right) \cdot \frac{n}{\theta^n} \left[\frac{1}{n+1} y^{n+1}\right]_0^\theta \\
&= \left(\frac{n+1}{n}\right) \cdot \frac{n}{\theta^n} \cdot \frac{\theta^{n+1}}{n+1} = \theta \\
E(\hat{\theta}_2) &= \theta
\end{aligned}
$$

$\hat{\theta}_2$ is an unbiased estimator of $\theta$.

$$
\begin{aligned}
Var(\hat{\theta}_2) &= E(\hat{\theta}_2^2) - [\underbrace{E(\hat{\theta}_1)}_{\theta}]^2 \\
E(\hat{\theta}_2^2) &= \int_0^\theta \left(\frac{n+1}{n}\right)^2 y^2 \delta_{Y_{(n)}} \, dy \\
&= \left(\frac{n+1}{n}\right)^2 \int_0^\theta y^2 \cdot n \cdot \frac{1}{\theta}\left(\frac{y}{\theta}\right)^{n-1} \, dy \\
&= \left(\frac{n+1}{n}\right)^2 \frac{n}{\theta^n} \int_0^\theta y^{n+1} \, dy \\
&= \left(\frac{n+1}{n}\right)^2 \cdot \frac{n}{\theta^n} \left[\frac{1}{n+2} y^{n+2}\right]_0^\theta \\
&= \left(\frac{n+1}{n}\right)^2 \frac{n}{\theta^n} \frac{\theta^{n+2}}{n+2} \\
&= \frac{(n+1)^2}{n(1+2)} \cdot \theta^2 \\
Var(\hat{\theta}_2^2) &= \frac{(n+1)^2}{n(n+1)} \cdot \theta^2 - \theta^2 \\
&= \theta^2 \left[\frac{(n+1)^2 - n(n+2)}{n(n+1)}\right]
\end{aligned}
$$

$$= \frac{\theta^2}{n(n+2)}$$

So what do we have?

$$\begin{aligned} Var(\hat{\theta}_1) &= \frac{\theta^2}{3n} \\ Var(\hat{\theta}_2) &= \frac{\theta^2}{n(n+2)} \\ eff(\hat{\theta}_1, \hat{\theta}_2) &= \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)} \\ &= \frac{\frac{\theta^2}{n(n+2)}}{\frac{\theta^2}{3n}} = \frac{3}{n+2} \\ \frac{3}{n+2} &\leq 1 \text{ if } n > 1, \frac{3}{n+2} < 1 \\ n \to \infty &\frac{3}{n+2} \to \infty \end{aligned}$$

So the meaning of this ratio is that as $n$ increases, the efficiency of procedure 1 decreases compared to the efficiency of procedure 2. It's definitely better to use the second procedure to estimate $\theta$.

---

## 7.2 Consistency

Minimum requirement an estimator should have. More rigorously:

**Def** We say $\hat{\theta}_n$ (subscript meaning it was based on $n$ observations) is a *consistent estimator* of $\theta$ if $\hat{\theta}_n \xrightarrow{P} \theta$ as $n \to \infty$.

Recall that converges with probability ($\xrightarrow{P}$) means:

$$\lim_{n \to \infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0, \forall \varepsilon > 0$$

i.e. the more and more elements we have, the more our estimates will converge to the actual value.

Recall:

$$\lim_{n\to\infty} a_n = a$$
$$\forall \varepsilon > 0, \exists N(\varepsilon) \implies |a_n - a| < \varepsilon \text{ if } n \geq N(\varepsilon)$$

What does this say? It means we can get as close as $a$ as we'd like by increasing $n$.

But this isn't the same with random variables, because they have **random fluctuations**! For example, if we had a fair coin, we'd try flipping many many times in order to try to induct that the probability is 0.5, but after many many trials, say 10000, we may only have 4820 heads or something of the like. i.e. we increased the sample size but we noticed an even bigger deviation and this happens in practice because of randomness. So what is the complement of converges with probability?

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1, \forall \varepsilon > 0$$

So this may not happen with certainty (because of random fluctuations, but the complement will converge to 1).

---

**Ex** $X_i \sim Ber(p)$

$$X_i = \begin{cases} 1 & p \\ 0 & 1-p \end{cases}, i = 1, \ldots, n.$$

What justifies us in using $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$? We said we were mimicking that proportion in the population. Now we want to formalize this. To establish that $\hat{p}_n$ is on the right track:

$$\lim_{n\to\infty} P(|\hat{p}_n - p| > \varepsilon) = 0, \forall \varepsilon > 0$$

What's our big hammer here? Chebychev's inequality. Can we use that here? Recall:

$$P(|X - E(x)| > k\sigma) \leq \frac{1}{k^2}$$

Our random variable $X$, is $\hat{p}_n$. The mean of $\hat{p}_n$ is $p$. Now $\varepsilon$ plays the role of our $k\sigma$, i.e.

$$E(\hat{p}_n) = p$$
$$\varepsilon = k\sqrt{var(\hat{p}_n)}$$
$$Var(\hat{p}_n) = \frac{p(1-p)}{n}$$

$$k = \frac{\varepsilon}{\sqrt{Var(\hat{p}_n)}}$$

So if we want to use Chebychev's:

$$\begin{aligned} P(|\hat{p}_n - p| > \varepsilon) &\leq \frac{1}{\left[\frac{\varepsilon}{\sqrt{Var(\hat{p}_n)}}\right]^2} \\ P(|\hat{p}_n - p| > \varepsilon) &\leq \frac{\left(\frac{p(1-p)}{n}\right)}{\varepsilon^2} \\ p(1-p) \leq \frac{1}{4} \implies &= \frac{p(1-p)}{n\varepsilon^2} \\ &\leq \frac{1}{4n\varepsilon^2} \to 0, \text{ as } n \to \infty, \forall \varepsilon > 0 \end{aligned}$$

So we have obtained:

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

So what if we want to decrease $\varepsilon$ as $n$ increases?

$$\begin{aligned} \varepsilon_n &= \frac{\log n}{\sqrt{n}} \\ P\left(|\hat{p}_n - p| > \frac{\log n}{\sqrt{n}}\right) &\leq \frac{1}{4n\left(\frac{\log n}{\sqrt{n}}\right)^2} \leq \frac{1}{4\log n} \to 0, \text{ as } n \to \infty \end{aligned}$$

So this tells us the rate at which we can decrease $\varepsilon$. This brings us closer to the definition of limits from Calculus from numbers.

---

So we can use Chebyshev's inequality to establish consistency for many estimators.

Suppose $X_1, \ldots, X_n$ have the same mean $\mu$ and variance $\sigma^2$. Suppose further that $Cov(X_i, X_j) = 0, i \neq j$. Then

$$\overline{X}_n \xrightarrow{P} \mu$$

which we can easily show using Chebyshev's inequality as well.

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

Recall that Chebyshev's:

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

So $\overline{X}_n \to X, \mu \to E(x), \varepsilon \to k\sqrt{Var(x)}$

$$\begin{aligned}
\varepsilon &= k\sqrt{Var(\overline{X}_n)} \\
k &= \frac{\varepsilon}{\sqrt{Var(\overline{X}_n)}} \\
Var(\overline{X}_n) &= \frac{\sigma^2}{n} \\
P(|\overline{X}_n - \mu| > \varepsilon) &\leq \frac{1}{\left(\frac{\varepsilon}{\sqrt{\frac{\sigma^2}{n}}}\right)^2} \\
&= \frac{\sigma^2}{n\varepsilon^2} \to 0, \text{ as } n \to \infty, \text{ as long as } \sigma^2 \text{ finite}
\end{aligned}$$

But can we also say:

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

No, there is much more difficulty with this, which we'll see next class.

# 8 02/06/18

## 8.1 Consistency

Last class we discussed consistency and went over a few examples, played a bit with $\varepsilon$ and convergent probability. We then extended the consistency result for sample proportions to sample averages for estimating the population average. Then we wanted to start discussing the consistency of population variance, although we noted that the usual big hammer (Chebyshev's/Markov's Inequality) would be harder to use.

$\hat{p}_n \xrightarrow{p} p \to p = P(X = 1)$

This essentially justifies using relative frequency as an estimate to $p$. As we repeat our

experiments more and more, the relative frequency $\hat{p}_n$ gets closer and closer to $p$. The whole theorem was trying to find an upper bound of the difference between relative frequency and $p$, using Chebyshev's Inequality (a special case of Markov's).

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

$$\begin{aligned} P(|\hat{p}_n - p| > \varepsilon) &= P(|\hat{p}_n - p|^2 > \varepsilon^2) \\ &\overset{\text{Markov's}}{\leq} \frac{E[|\hat{p}_n - p|^2]}{\varepsilon^2} \end{aligned} \tag{5}$$

Recall: If $g$ is a non-negative function and $X$ a r.v., then

$$P(g(X) \geq \lambda) \leq \frac{E[g(X)]}{\lambda}, \lambda > 0$$

To apply Markov's inequality to obtain (5), $g(x) = (x - E(x))^2$. Now replace $X$ by $\hat{p}_n$. Note that $E(\hat{p}_n) = p$. Why did we square before applying Markov's? Because it is easier to compute variance with.

$$\begin{aligned} E[(\hat{p}_n - p)^2] &= Var(\hat{p}_n) = \frac{p(1-p)}{n} \\ p(|\hat{p}_n - p| > \varepsilon) &= p(|\hat{p}_n - p|^2 > \varepsilon^2) \\ &\leq \frac{E[|\hat{p}_n - p|^2]}{\varepsilon^2} = \frac{p(1-p)}{n\varepsilon^2} \to 0, \text{ as } n \to \infty, \forall \varepsilon > 0 \end{aligned}$$

This means the variance approaches 0, so the estimator gets closer and closer to its center, what its trying to estimate.

**Theorem** (Thm 9.1, p450)
Suppose $\hat{\theta}_n$ is an estimate for $\theta$ and $MSE = (\hat{\theta}) \to 0$ as $n \to \infty$. Then $\hat{\theta}_n \xrightarrow{P} \theta$ meaning that $\hat{\theta}_n$ is a consistent estimator of $e$.

**Proof**

$$\begin{aligned} P(\left|\hat{\theta}_n - \theta\right| > \varepsilon) &= P(\left|\hat{\theta}_n - \theta\right|^2 > \varepsilon) \\ &\overset{\text{Markov's}}{\leq} \frac{E[\left|\hat{\theta}_n - \theta\right|^2]}{\varepsilon^2} \\ &= \frac{MSE(\hat{\theta}_n)}{\varepsilon^2} \to 0 \end{aligned}$$

So $\hat{\theta}_n \xrightarrow{P} \theta$.

**Corollary** Suppose $\hat{\theta}_n$ is an unbiased estimator of $\theta$ such that $Var(\hat{\theta}_n) \to 0$ as $n \to \infty$. Then $\hat{\theta}_n \xrightarrow{P} \theta$, i.e. $\hat{\theta}_n$ is a consistent estimator of $\theta$.

**Recall** Proportion:

$$Var(\hat{p}_n) = \frac{p(1-p)}{n} \to 0, \text{ as } n \to \infty$$

Sample average:

$$Var(\overline{X}_n) = \frac{\sigma_X^2}{n} \to 0, \text{ as } n \to \infty$$

$$\begin{aligned} E(\hat{p}_n) &= p \\ E(\overline{X}_n) &= \mu_X \\ MSE(\hat{p}_n) &= Var(\hat{p}_n) \\ MSE(\overline{X}_n) &= Var(\overline{X}_n) \end{aligned}$$

**Recall**

$$MSE(\hat{\theta}_n) = Var(\hat{\theta}_n) + \underbrace{Bias^2(\hat{\theta}_n)}_{(E(\hat{\theta}_n)-\theta)^2} \tag{6}$$

**Corollary 2** Suppose $\hat{\theta}_n$ is asymptotically unbiased for $\theta$, i.e. $\lim_{n\to\infty} E(\hat{\theta}_n) = \theta$, and $Var(\hat{\theta}_n) \to 0$ as $n \to \infty$, then $\hat{\theta}_n \xrightarrow{P} \theta$, i.e. $\hat{\theta}_n$ i a consistent estimator of $\theta$.

Using (6):

$$\lim_{n\to\infty} MSE(\hat{\theta}_n) = \lim_{n\to\infty} Var(\hat{\theta}_n) + \underbrace{\lim_{n\to\infty} Bias^2(\hat{\theta}_n)}_{0} = \lim_{n\to\infty} Var(\hat{\theta}_n) = 0$$

---

Suppose $X_1, \ldots, X_n$ for a random sample (i.e. independent and identically distributed) from a population with mean $\mu$ and variance $\sigma^2$. We want to estimate $\sigma^2$. We learned that perhaps a reasonable estimator is:

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

$$E(S^2) = \sigma^2$$

But we need to divide by $n-1$.

$$S^2_{n1*} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \left(1 - \frac{1}{n}\right) S^2_n$$
$$E(S^2_{n,*}) = \left(1 - \frac{1}{n}\right) E(S^2_n) = \left(1 - \frac{1}{n}\right) \sigma^2$$

i.e. $S^2_{n,*}$ is biased, but it will be our maximum likelihood estimator later.

Can we show:

$$S^2_n \xrightarrow{P} \sigma^2?$$

$$\begin{aligned} P(\left|S^2_n - \sigma^2\right| > \varepsilon) &= P(\left|S^2_n - \sigma^2\right| > \varepsilon) \\ &\leq \frac{E[(S^2_n - \sigma^2)]}{\varepsilon^2} \\ &= \frac{Var(S^2_n)}{\varepsilon^2} \end{aligned}$$

Note that all of these depend heavily on the second moment and give us difficulty. They require variance, but Kolmogarov's Theorem doesn't and makes it important/hard to prove.

## 8.2 Kolmogarov's Theorem (Law of Large Numbers)

Suppose $X_1, \ldots, X_n$ is a random sample with population mean $\mu$. Then $\overline{X}_n \xrightarrow{P} \mu = E(X)$.

$$\begin{aligned} S^2_{n,*} &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \\ &= \underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i^2}_{E(X^2)} - \overline{X}^2_n \end{aligned}$$

Now $X_1, \ldots, X_n$ *iid* and $X^k_1, \ldots, X^k_n$ *iid* $\to E(X^k)$

$$\frac{1}{n}\sum_{i=1}^{n}\underbrace{X_i^k}_{Y_i} \to E(X^k)$$
$$\frac{1}{n}\sum_{i=1}^{n} X_i^k = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}_n$$

Using Kolmogarov's theorem:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \to E(X^2) \tag{7}$$

$$\overline{X}_n \xrightarrow{P} E(X) \tag{8}$$

$$(?)\overline{X}_i^2 \to [E(X)]^2 \tag{9}$$

---

Thm 9.7, page 451

If $\hat{\theta}_n \xrightarrow{P} \theta$, then $\hat{\phi}_n \xrightarrow{P} \phi$. Then

a) $\hat{\theta}_n\hat{\phi}_n \to \theta\phi$

b) $\hat{\theta}_n \pm \hat{\phi}_n \to \theta \pm \phi$

c) $\frac{\hat{\theta}_n}{\hat{\phi}_n} \to \frac{\theta}{\phi}$, provided that $\hat{\phi}_n \neq 0$ and $\phi \neq 0$.

d) $g(\hat{\theta}_n) \to g(\theta)$ if $g$ is a continuous function.

This is called the continuous mapping theorem.

Then (9) follows from $9.2(d)$, the continuous mapping theorem.

$$\hat{\theta}_n \xrightarrow{P} \theta : P(\left|\hat{\theta}_n - \theta\right| > \varepsilon) \to 0, ash \to \infty, \forall \varepsilon > 0$$

Step (8): $\overline{X}_n^2 \to [E(X)]^2$

Step (9): Using $9.2(b)$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2 \to \underbrace{E(X^2) - [E(X)]^2}_{Var(X)=\sigma^2}$$

So $S_{n,*}^2 \xrightarrow{P} \sigma^2$

Question:

$$S_{n,*}^2 = \frac{1}{n}\sum_{i=1}^{n} \underbrace{(X_i - \overline{X}_n)^2}_{Z_i} = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

Why can't we use Kolmogarov's theorem directly here? They are not independent.

$$Z_i = (X_i - \overline{X}_n)^2$$

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

Consistency means being right-headed, if you have an estimator that is not consistent, throw it out.

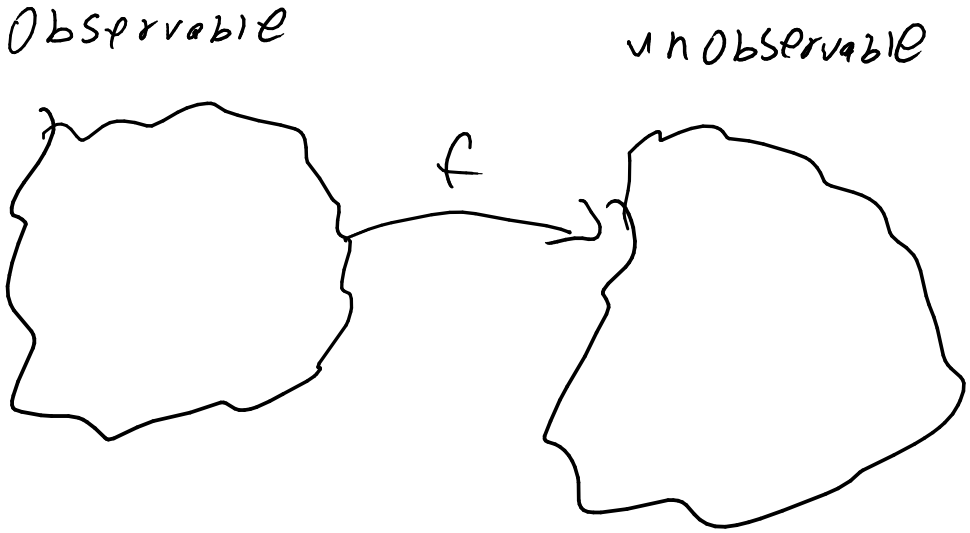
Now suppose that you get the whole population (magically somehow), can you hit the target parameter? This is the minimum requirement, consistency.

Recall that before our goal was to get the "best" estimator, the one that is closest to the target. Sufficiency paves the way to that.

## 8.3 Sufficiency

Sufficiency was first introduced by Fisher, at first for compression.

**Likelihood Inference**




For example: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, where $x$ is observable and $\mu$ and $\sigma$ are unobservable.

Lets say that $\sigma$ is known and $\mu$ is unknown. How do we quantify this information? What will compressing the data do? If we compress, we also want to be able to decompress to get the original data. It's a bit too much to expect to get the exact original data after decompressing, so maybe we can lower our expectations.

$$\vec{x} = (X_1, \ldots, X_n)$$
$$\vec{y} = (Y_1, \ldots, Y_n)$$

Now suppose $T$ is the compressed value as follows:

$$T = T(X_1, \ldots, X_n)$$
$$T : \mathbb{R}^n \to \mathbb{R}^m$$

We want $m < n$, or else we aren't compressing anything. We want to work with a smaller dimension, the smaller the $m$ the better.

We say that $\vec{X}$ & $\vec{Y}$ are T-similar if

$$P_{\vec{X}|T}(\vec{u}|t_1\theta) = P_{\vec{Y}|T}(\vec{u}|t_1\theta), \forall \vec{u}$$

Given compressed values, we want to come back to something similar (T-similar) if not the same. They are similar, we cannot tell the difference, we come up with something equally likely.

A realization of $\vec{X}$ and a realization of $\vec{Y}$ are T-similar if

a) $\vec{X}$ & $\vec{Y}$ are T-Similar

b) $T(\vec{X}) = T(\vec{Y})$