MATH 324: Statistics

Julian Lore

Last updated: January 11, 2018

Notes from Masoud Asgharian's Winter 2018 lectures.

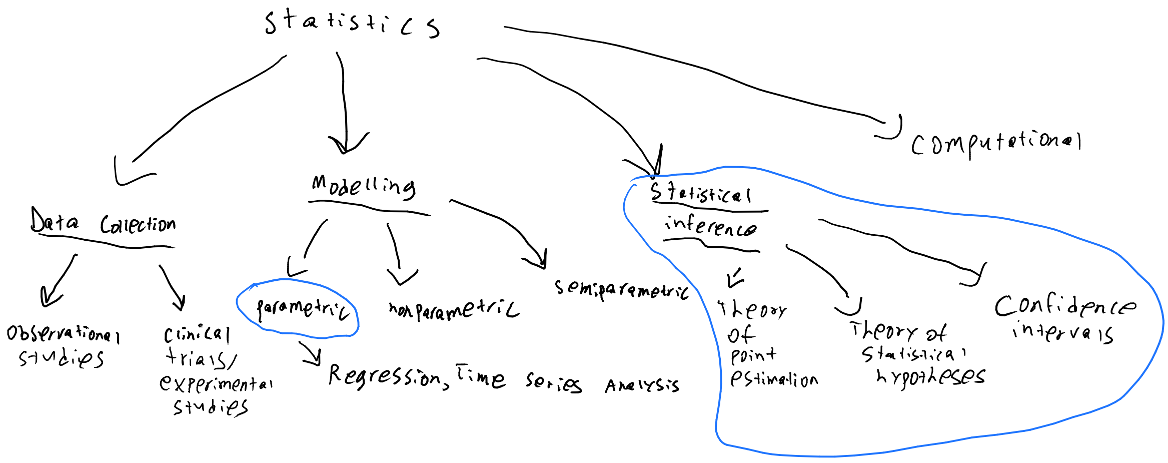
Contents

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. $\to X|\mu, \sigma^2 \sim N(\mu, \sigma^2)$, 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: $x_1, \ldots, x_n, \mu \rightarrow$ 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?

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

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:

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

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?

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

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

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

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.