Basic Stats review

Estimators: A function $\hat{\theta}$ of the random sample of data we use to estimate the value of some unknown population parameter θ

- Think of $\hat{\theta}$ as our "best guess" for θ
- There can be many estimators for any given parameter. We want to find and use the "best" one (high acc, low variance) or (low bias, low variance)

Expected Value: Basically the average of a random variable.

(continuous RV)

$$E[X] = \int_{-\infty}^{\infty} x f_\theta(x)\, dx \quad , \quad X \sim f_\theta(x)$$

$f_\theta(x)$ ↓ PDF

$$\int_{-\infty}^{\infty} f_\theta(x)\, dx = 1$$

(discrete RV)

$$E[X] = \sum_{x \in \mathcal{X}} x\, p(x) \quad , \quad p(x) = P(X = x)$$

$p(x)$ ↳ PMF

$$\sum_{x \in \mathcal{X}} p(x) = 1 \quad , \quad \mathcal{X} := \text{range of } X$$

Variance: Measures deviations from the expected value. It gives us an idea of how the observations are distributed.

- We want as little variance as possible. Squaring the differences makes values bigger and overestimates on purpose so that models learn better.

- But, paradoxically, we don't want 0 variance. Anyone knows why? Think back to Econ 57.

(continuous RV)

$$Var(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 f_\theta(x)\, dx$$

(Discrete RV)

$$Var(X) = \sum_{x \in X} [x - E(X)]^2 p(x)$$

We always seek to minimize the variance but can you tell me the intuition behind squaring the differences?

- Overestimate on purpose so models learn better.
- Ignore negative values since we only care about the distance to the expectation.

Ex 1.

Suppose $X \sim U[a,b]$. Then,

$$f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

$$E(X) = \int_a^b x\, f(x)\, dx = \int_a^b x \frac{1}{b-a}\, dx$$

$$= \frac{1}{b-a} \int_a^b x\, dx = \frac{x^2}{2(b-a)} \Big|_a^b$$

$$= \frac{b^2}{2(b-a)} - \frac{a^2}{2(b-a)} = \frac{b^2-a^2}{2(b-a)}$$

$$= \frac{(\cancel{b-a})(b+a)}{2(\cancel{b-a})} = \frac{b+a}{2}$$

$$\Rightarrow E[X] = \frac{b+a}{2}$$

Ex 2.

There is an alternative specification for the variance. Useful if we know the expected value in advance

the expected value in advance.

Suppose $\mu = E(X)$

$$\begin{aligned}
Var(X) &= E\left[(X-\mu)^2\right] \\
&= E\left[X^2 - 2X\mu + \mu^2\right] \\
&= E\left[X^2\right] - E\left[2X\mu\right] + E\left[\mu^2\right] \\
&= E\left[X^2\right] - 2\mu E\left[X\right] + \mu^2 \\
&= E\left[X^2\right] - 2\mu^2 + \mu^2 \\
&= E\left[X^2\right] - \mu^2
\end{aligned}$$

Covariance: Measures how two

variables move (linearly) together. In other words, it gives us information about the joint variability of two RVs

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

*

$$E[XY] = \sum (x \cdot y) \cdot p(x,y)$$

↳ joint prob.

or

$$E[XY] = \int (x \cdot y) f_\sigma(x,y)\, dx$$

Using the fact that the Expected value is

a

linear operator and that μ_x, μ_y are deterministic

Prove that $E[(X-\mu_x)(Y-\mu_y)] = E[XY] - \mu_x \mu_y$

$$E[XY - X\mu_y - \mu_x Y + \mu_x \mu_y]$$

$$= E[XY] - \mu_y E[X] - \mu_x E[Y] + E[\mu_x \mu_y]$$

$$= E[XY] - \mu_y \mu_x - \mu_x \mu_y - \mu_x \mu_y$$

$$= E[XY] - \mu_y \mu_x$$

" Higher values of one variable tend to correspond in higher values of the other one"

- Directionality goes both ways, be careful.
- Can be generalized to multiple variables with variance-covariance matrix

Properties of $E(X)$, $Var(X)$, $Cov(X,Y)$

Let X, Y, W, Z be RV and $a, b, c, d \in \mathbb{R}$

1) $E(aX + bY + c) = aE(X) + bE(Y) + c$

↳ Linear operator

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

Proof.

$$Var(b+aX) = E\left[(a + bX - E(a+bX))^2\right]$$

$$= E\left[(a + bX - a - bE(X))^2\right]$$

$$= E\left[(bX - bE(X))^2\right]$$

$$= E\left[(b(X - E(X)))^2\right]$$

$$= E\left[b^2 (X - E(X))^2\right]$$

$$= b^2 E\left[(X - E(X))^2\right]$$

$$= b^2 Var(X)$$

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

* Hint: start with

$$Var(\cdot) = E\left[(aX \pm bY - E(aX \pm bY))^2\right]$$

$$4)\ Cov(aX + bY, cW + dZ) =$$

$$a \cdot c \cdot Cov(X, W) + ad\ Cov(X, Z) +$$

$$bc\ Cov(Y, W) + bd\ Cov(Y, Z)$$

Now, these are theorethical operations. Which means that we assume we know the underlying population distribution. In practice, we cannot know this. So adjustments must be made to correct for over- or under-estimations. This is why

population and sample operations differ slightly. A correction you've certainly encountered is dividing a variance by $N-1$ rather than N. This is called Bessel's correction. It addresses the "degrees of freedom" in our sample estimators.

The idea is the following. To compute the variance we first need to estimate the average. By doing so we lose one degree of freedom. That is, we add a constraint to the estimation that restricts how the data can vary.

Exc. Give me three numbers that sum to 10. Notice how you only really

have two genuine choices. The third one is fully determined. That is, it is mathematically determined by our constraint the these must sum to 0.

Same idea applies for the sample variance Suppose we have a sample $\{2,5,8\}$. Its mean is 5. Substracting each value from the mean gives us our deviations $\{3, 0, -3\}$. Notice they sum to 0. This is precisely our constraint.

Proof:

$$\bar{x} = \frac{1}{n}\sum x_i$$

$$\sum (x_i - \bar{x}) = \sum x_i - \sum \bar{x} = \sum x_i - n\bar{x}$$

$$= \sum x_i - n\frac{1}{n}\sum x_i = 0.$$

Deviations from the mean MUST sum to 0 by definition of the sample mean.

We can divide the sample mean by n because all we need to estimate it is the data itself, without further constraints

So, once we know $n-1$ deviations the n^{th} one is fully determined. It must be whatever number makes the sum 0.

Had we known the true population mean (μ) rather than having to estimate it with $\bar{x}$, we wouldn't have lost this df and no correction would be needed.

Using $n-1$ rather than n gives us

an
unbiased estimator of the true population variance. This adjustment compensates for the constraint imposed by having to estimate the mean.

In general, the more parameters you need to estimate, the more degrees of freedom you lose. This has implications for inference, specifically data requirements. To estimate a variance you need at least 2 data points. In a simple regression, you need at least three. Can you see why? We need to estimate two coefficients (β_0, β_1). What if I have a regression with k vars? We need $n > k+1$ data points.

This is a fundamental insight of statistical inference directly linked to statistical power, sample sizes, precision, and model selection.
Now you know why a neural network, which has millions of parameters, needs so much data to work.

We care about these corrections because
we want the best estimators.
Unbiasedness and Efficiency are the two properties we will look, and the focus of the next section. Now, let's look at some code to build an intuition.

Small sample properties

Usually, specially in fields like macro, we must deal with small samples (a couple of hundreds data points).

- Recall as a general rule of thumb that, in order to have at least a little confidence in your estimation / inference, you need at least 30 to 50 data points. The more the better.

In small samples we want to pick the estimators that are unbiased and efficient

Unbiasedness: A estimator $\hat{\theta}$ is unbiased for a population parameter θ if $E[\hat{\theta}] = \theta$, $\forall \theta$

Efficiency: Many estimators could be unbiased. In that case, we want the one with lower variance

(iid)

Exc. Let $X \sim f(\mu, \sigma^2)$. Prove the sample mean is an unbiased estimator for μ.

Compute the Variance of the sample mean.

Large Sample properties

Moreover, as the sample size increases, there are additional properties we want to see.

Let θ be the population parameter and $\hat{\theta}_N$ the estimator with a sample of size N.

Ideally, as $N \to \infty$, the probab. that $\hat{\theta}_N$ and θ become trivially close together is 1. That is, a good estimator will converge to

the true population value if we have a lot of data.

Convergence in Probability:

$\hat{\theta}_N$ converges in probability to θ if, for any arbitrary $\varepsilon > 0$,

$$\lim_{N \to \infty} \text{Prob}\left(|\hat{\theta}_N - \theta| < \varepsilon\right) = 1$$

or

$$\lim_{N \to \infty} \text{Prob}\left(|\hat{\theta}_N - \theta| > \varepsilon\right) = 0$$

If any of these hold, then we denote $\text{plim}(\hat{\theta}_N) = \theta$ or $\hat{\theta}_N \xrightarrow{p} \theta$ or $\hat{\theta}_N = \theta + o_p(1)$

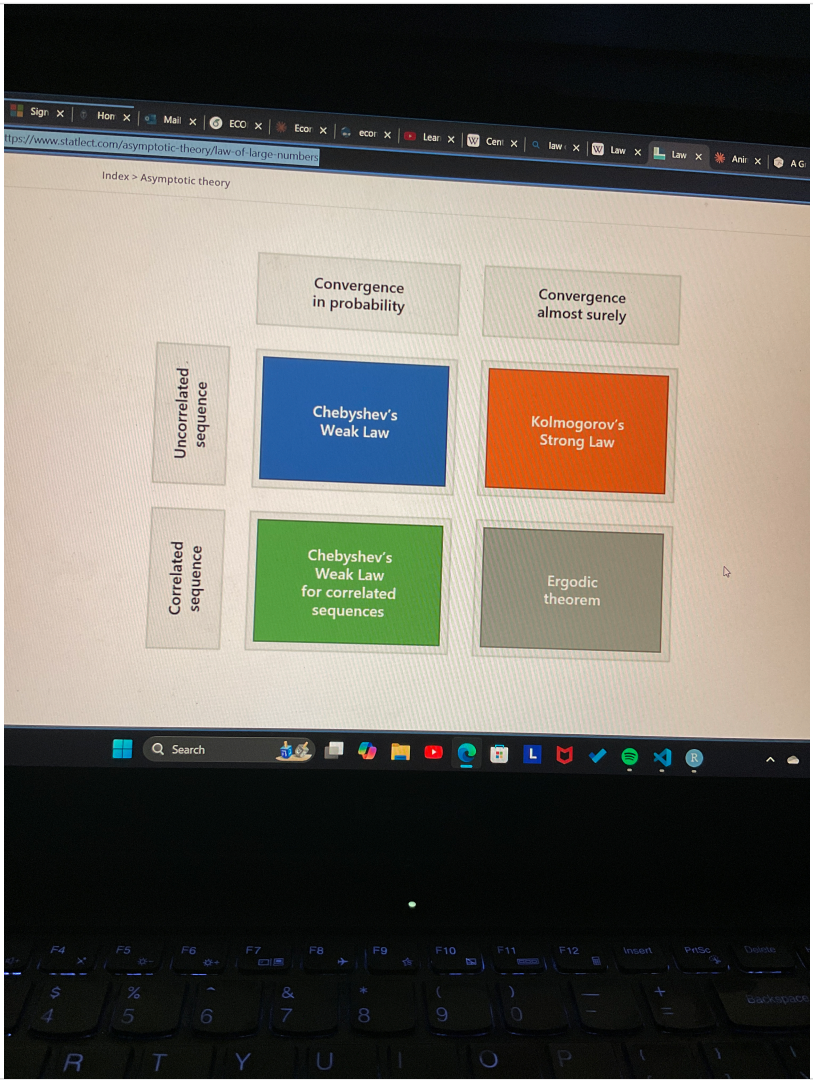
↳ converges to 0

Consistency: If the estimator $\hat{\theta}_n$ converges in probability to θ then we say it is consistent for θ

Law(s) of Large Numbers (LLN)

A Law of Large Numbers is a proposition that provides a set of sufficient conditions for the convergence of the sample mean to a constant (usually the expected value of the distribution from which the sample was drawn)

There are a bunch of LLNs, but they are broadly classified into weak (converge in probability) and strong (converge almost surely).



LLNs are important because they guarantee stable long-term results for the averages of some random events.

But they do not apply to all distributions, like Cauchy or Pareto which have heavy tails that make it hard or impossible to find an expectation.

In Economics, a relevant limitation is Selection Bias. Briefly, when we run an experiment we aim to select participants at random. Most of the times the

incentives to participate causes certain types of participants to self-select, biasing the results. Unfortunately, simply recruiting more people or running more trials won't help in reducing the bias. Unless you recruit everyone, at which point you face a budget constraint.

Central Limit Theorem (CLT)

So, the LLN says that the average of a sample of iid obs. will converge to the expected value of the distribution it was drawn from.

The CLT goes a step further to claim that if we normalize the sample mean then the resulting distribution will converge to the standard Normal Distribution. Even if the original variables themselves are not normally distributed!

I want to emphasize that the distribution of averages is what converges to a standard normal $N(0,1)$ and not the sample itself

CLT normally requires iid

observations but, like with LLNs, these can be relaxed under certain conditions.

(Lindeberg-Levy CLT)

Suppose $X_1, X_2, X_3, \ldots$ are iid random variables with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2 < \infty$. Then as $n \to \infty$, the RVs $\sqrt{n}(\bar{X}_n - \mu)$ converges in distribution to $\mathcal{N}(0, \sigma^2)$

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

Recall that since X_i is a RV, each has its own distribution and hence expectation. $E(X_i) = \mu$ assumes that all of them have the same expectation, μ.

In the case $\sigma > 0$, convergence in distribution means that the cumulative probability function (CDF) of $\sqrt{n}(\bar{X}_n - \mu)$ converge pointwise to the CDF of the $N(0, \sigma^2)$ distribution.

For every $z \in \mathbb{R}$

Standardization

$$\lim_{n \to \infty} P\left[\sqrt{n}\,(\bar{x}_n - \mu) \le z\right]$$

$$= \lim_{n \to \infty} P\left[\frac{\sqrt{n}\,(\bar{x}_n - \mu)}{\sigma} \le \frac{z}{\sigma}\right]$$

$$= \Phi\left(\frac{z}{\sigma}\right)$$