1 Review of Probability

1.1 Basics

We will start by briefly reviewing some basic definitions from probability.

Definition 1.1. Random Variable (RV). A random variable is a function

$$X : \Omega \to \mathbb{R}$$

that assigns to each outcome ω (in the sample space Ω) a real number $X(\omega)$.

Definition 1.2. Probability. Given an RV X and a set $A \subset \mathbb{R}$, let $X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$. Define the probability that X is in A to be

$$\mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\}).$$

The CDF essentially contains all of the information about an RV.

Definition 1.3. Cumulative Distribution Function (CDF). For a RV X, define the CDF as the function $F_\theta : \mathbb{R} \to [0, 1]$ with

$$F_\theta(x) = \mathbb{P}(X \leq x),$$

where θ is a parameter (i.e., fixed/deterministic/nonrandom).

Theorem 1.1. Properties of CDF. A function $F_\theta : \mathbb{R} \to [0, 1]$ is a CDF for some probability $\mathbb{P}$ if and only if F satisfies the following three conditions:

(a) F is non-decreasing: $x_1 < x_2$ implies that $F(x_1) \leq F(x_2)$

(b) F is "normalized":

$$\lim_{x\to-\infty} F(x) = 0$$

(c) F is right-continuous:

$$\lim_{x\to y^+} F(x) = F(y).$$

Theorem 1.2. Uniqueness of a CDF. Let X be an RV with CDF F, Y an RV with CDF G, and suppose $F(x) = G(x)$ for all x. Then $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ for all A.

Definition 1.4. Discrete RV & Probability Mass Function (PMF). A discrete RV X is a RV that takes at most countably many values x. That is, X takes on either finitely many values (e.g., $x \in \{0, 1, 2, 3\}$) or countably many values ($x \in \{x_1, x_2, \ldots, x_n, \ldots\}$). We define the PMF for X as

$$f_\theta(x) = \mathbb{P}(X = x).$$

Definition 1.5. Continuous RV & Probability Density Function (PDF). A continuous RV X is a RV with uncountably many values (i.e., values that are on the real line or a subset of the real line). Specifically, it is defined as an RV for which there exists a function f_θ satisfying:

- $f_\theta(x) \geq 0$ for all x
- $\int_{-\infty}^{\infty} f_\theta(x)dx = 1$
- for any $a, b \in \mathbb{R}$ with $a < b$,

$$\mathbb{P}(a < X < b) = \int_a^b f_\theta(x)dx.$$

The PDF is defined as that function f_θ.

Note that if X is a continuous RV, then the probability that it takes any one value is 0, i.e.,

$$\mathbb{P}(X = x) = 0$$

for any x in the support.

For a discrete RV, the relationship between the PMF and CDF is as follows:

$$\begin{aligned} F_\theta(x) &= \sum_{k \leq x} f_\theta(k) \\ &= \sum_{k \leq x} \mathbb{P}(X = k). \end{aligned}$$

Thus, like the CDF, the PMF essentially contains all the information about a discrete RV.

Similarly, for a continuous RV, the relationship between the PDF and CDF is that

$$F_\theta(x) = \int_{-\infty}^{x} f_\theta(t)dt.$$

This follows by a Fundamental Theorem of Calculus, but it also makes sense that we would integrate for the continuous RV, since we summed for the discrete RV and integration is a generalization of summation. Additionally, by the other Fundamental Theorem of Calculus, we have that

$$f_\theta(x) = F'_\theta(x)$$

at all points x for which F_θ is differentiable.

1.2 Mean & Variance of RVs

In this course, we will typically write PDFs, CDFs, PMFs, etc. with the parameter(s) in the subscript. For example, we wrote $F_\theta(x)$ above for the CDF. The reason is that it emphasizes that, while the parameter θ is a fixed value, the function does ultimately depend on that value. You may have seen other notations before though, such as where the RV is in the subscript (e.g., $F_X(x)$ or the parameter is being conditioned on (e.g., $F(x|\theta)$. Also keep in mind that we can use θ as a vector containing multiple parameters, e.g., $\theta = (\mu, \sigma^2)'$. [Sometimes I might drop the θ subscript though, if it's clear from context.)

However, the parameter θ is not necessarily the only quantity we would care about for a random variable. Commonly, we would also care about the mean and variance, for example.

Definition 1.6. Mean of an RV. The mean or expectation of an RV X is defined as

$$\mathbb{E}_\theta(X) = \int_{-\infty}^{\infty} t f_\theta(t) dt$$

if X is continuous and

$$\mathbb{E}_\theta(X) = \sum_{k \in K} k f_\theta(k)$$

if X is discrete with support K. In either case, $\mathbb{E}_\theta(X)$ is often denoted as μ.

What if instead of having X, we have some function of X (which is thus also an RV)? How can we compute its expectation? LOTUS!

Theorem 1.3. Law of the Unconscious Statistician (LOTUS). Let X be an RV with CDF F_θ, and suppose we have some function $g(X)$. The expectation of $g(X)$ can be calculated as

$$\mathbb{E}g(X) = \int_{-\infty}^{\infty} g(t) f_\theta(t) dt$$

if X is continuous and

$$\mathbb{E}g(X) = \sum_{k \in K} g(k) f_\theta(k)$$

if X is discrete with support K, under some assumptions on g.

Definition 1.7. Variance of an RV. The variance of an RV X is

$$\mathrm{Var}_{\theta_X} = \mathbb{E}\left((X - \mathbb{E}X)^2\right).$$

By LOTUS, we can calculate the variance directly from the its definition as follows:

$$\mathrm{Var}_\theta(X) = \int_{-\infty}^{\infty} (t - \mathbb{E}(X))^2 f_\theta(t) dt$$

if X is continuous and

$$\text{Var}_\theta(X) = \sum_{k \in K} (k - \mathbb{E}(X))^2 f_\theta(k)$$

if X is discrete with support K.

A few useful facts for expectations and variances: Let $X_1, X_2, \ldots, X_p$ be RVs and $a_1, a_2, \ldots, a_p$ be scalars. Then we have

- Linearity of Expectation:

$$\mathbb{E}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i \mathbb{E} X_i$$

.

- Property of Variance for Linear Combinations:

$$\text{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \text{Var}(X_i) + 2 \sum_{i<j} a_i a_j \text{Cov}(X_i, X_j).$$

.

- Property of Variance:

$$\text{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2.$$

Definition 1.8. k-th (Central) Moment. The k-th (central) moment of an RV is defined as $\mathbb{E}(X^k)$, i.e., the expectation of the random variable raised to the k-th power, where k is a positive integer. [The k-th non-central moment about c refers to $\mathbb{E}[(X - c)^n]$, where c is a scalar. Unless otherwise specified, when we say “moment” we will be referring to the central moment, i.e., $c = 0$.]

Moments besides the first two (the mean and variance) can be helpful for describing different aspects of a distribution. For example, the third moment is the skewness, and the fourth moment is the kurtosis (which is related to the peakedness of a distribution at its center vs. the thickness of its tails).

Moments also are used for certain methods in statistics, such as the Method of Moments (MoM) approach to estimation.

Moments are also necessary for defining the moment-generating function (MGF) of a distribution.

Definition 1.9. Moment-generating Function (MGF). Let X be an RV with CDF F_θ. The MGF is

$$M_\theta(t) = \mathbb{E}(e^{tX})$$

The MGF is useful for several main reasons. Firstly, like the CDF, it completely characterizes an RV and is unique. This means that it can be a powerful tool for determining the distribution of a RV or transformation of RV. One of the exercises at the end of these notes shows an example of this. Secondly, if you know the MGF, then you can use it to easily calculate any moment you want, since

$$\mathbb{E}(X^k) = \frac{d}{dt} M_\theta(t)|_{t=0}.$$

How can we actually compute an MGF (or a moment, if we didn't know the MGF)? Simply apply LOTUS!

Note that the mean, variance, other moments, and MGF are all considered parameters, since they are non-random, fixed quantities. In addition to these parameters, what are some other parameters we might care about? There are many, many possibilities, depending on the particular problem, such as the mode, the interquartile range, the minimum, the maximum, etc. We will see so many examples in this class!

1.3 Examples of RVs

Here is a list of common distributions for RVs:

Discrete: Discrete Uniform, Bernoulli, Binomial, Negative Binomial, Geometric, Hypergeometric, Poisson

Continuous: Continuous Uniform, Normal, Exponential, Gamma, χ^2, Beta, t, F, Cauchy

You don't need to memorize the formulas for these — just look up the PDFs/PMFs, means, variance, etc. when you need them. But you should be comfortable knowing/figuring out which distribution to use for different situations (e.g., that the Poisson is often used for count data, that the normal distribution is bell-shaped symmetric, etc.), as well as being able to spot that you might need to use special properties of the RV to solve the problem (e.g., that the sum of iid exponential RVs is Gamma).

1.4 Bivariate Distributions

So far, we have only defined univariate proability distributions. Let us now define bivariate probability distributions, which are quite straightforward extensions of the univariate case.

Definition 1.10. Bivariate Cumulative Densitity Function (CDF). Let X and Y be RVs. Then the joint CDF is defined as

$$F_{\theta_X,\theta_Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y).$$

Definition 1.11. Bivariate Joint Probability Mass Function (PMF). Let X and Y be discrete RVs. The joint PMF for (X,Y) is defined as

$$f_{\theta_X,\theta_Y}(x,y) = \mathbb{P}(X = x, Y = y).$$

Definition 1.12. Bivariate Joint Probability Density Function (PDF). Let X be a continuous RV. A function $f_{\theta_X,\theta_Y}(x,y)$ is a joint PDF for (X,Y) if

(a) $f_{\theta_X,\theta_Y}(x,y) \geq 0$ for all (x,y)

(b) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\theta_X,\theta_Y}(x,y)dxdy = 1$

(c) for any set $A \subset \mathbb{R} \times \mathbb{R}$,

$$\mathbb{P}((X,Y) \in A) = \int\int_A f_{\theta_X,\theta_Y}(x,y)dxdy.$$

Even if we are interested in two variables together (and thus working with the joint distribution), we might also still care about the univariate distribution for one or both variables. We refer to such univariate distributions as the marginal distributions.

Definition 1.13. Marginal Distributions. If (X,Y) has joint distribution with PDF $f_{\theta_X,\theta_Y}(x,y)$, then the marginal distribution of X is

$$f_X(x) = \int_{-\infty}^{\infty} f_{\theta_X,\theta_Y}(x,y)dy$$

if X and Y are continuous and

$$f_X(x) = \mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y f_{\theta_X,\theta_Y}(x,y) \tag{1}$$

if X and Y are discrete. The marginal distributions are similarly defined for Y.

That is, we obtain the marginal distribution for X from the joint distribution for (X,Y) by summing or integrating out Y. It's like we are isolating X from (X,Y) by looking at the distribution that results from considering any and all possible values of Y.

But in general, while the marginal distributions might be helpful to look at, they are not sufficient for understanding the behavior of two variables – we need the joint distribution. The exception is if the two variables are independent, because then their joint distribution factors into the marginal distributions.

Definition 1.14. Two RVs X and Y are independent if for every set A and B,

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

It would be cumbersome in general to use the above definition to check if two RVs are independent, as we would have to check the probabilities for every set. But luckily, it suffices to check the PDF or PMF:

Theorem 1.4. Let X and Y have joint PDF $f_{\theta_X,\theta_Y}(x, y)$. Then X and Y are independent if and only if

$$f_{\theta_X,\theta_Y}(x, y) = f_{\theta_X}(x)f_{\theta_Y}(y).$$

If X and Y are independent, then for their expectations we have

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

One way of measuring dependence among two RVs is through the covariance.

Definition 1.15. Covariance. The covariance between X and Y is defined as

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)].$$

Note that the covariance is not everything! It does not in general completely characterize the joint distribution of two variables. Similar to how we have the mean and variance but ALSO higher-order moments (which all convey different information for an RV), we can have higher-order dependencies beyond the covariance.

Beware: if X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$. But if $\mathrm{Cov}(X, Y) = 0$, X and Y are not necessarily independent! There could be higher-order dependencies or that the covariance has “canceled out”!

Similarly to writing the variance as the difference of the second moment and the squared mean, for the covariance we can write

$$\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

In additional to joint and marginal distributions, we can define conditional distributions.

Definition 1.16. Conditional Probability Mass Function (PMF). The conditional PMF for X given Y is

$$f_{\theta_X,\theta_Y}(x|y) = \frac{\mathbb{P}(X=x,Y=y)}{\mathbb{P}(Y=y)} = \frac{f_{\theta_X}(x,y)}{f_{\theta_Y}(y)},$$

assuming $f_{\theta_Y}(y) > 0$. It is defined similarly for Y given X.

Definition 1.17. Conditional Probability Density Function (PDF). The conditional PDF for X given Y is

$$f_{\theta_X,\theta_Y}(x|y) = \frac{f_{\theta_X,\theta_Y}(x,y)}{f_{\theta_Y}(y)},$$

assuming that $f_{\theta_Y}(y) > 0$.

If we then want to calculate the conditional probability of event A given $Y = y$, then we simply use the conditional PDF:

$$\mathbb{P}(X \in A|Y = y) = \int_A f_{\theta_X,\theta_Y}(x|y)dx.$$

A very helpful theorem for handling conditional probabilities is as follows.

Theorem 1.5. Bayes' Theorem.

$$f_{\theta_X,\theta_Y}(y|x) = \frac{f_{\theta_X}(x|y)f_{\theta_Y}(y)}{f_{\theta_X}(x)}.$$

We can extend the definition of expectation to conditional expectation and variance.

Definition 1.18. Conditional Expectation & Variance The conditional expectation for X given Y is

$$\mathbb{E}_{\theta_X,\theta_Y}(X|Y=y) = \int_{-\infty}^{\infty} x f_{\theta_X,\theta_Y}(x|y)dx$$

if X and Y are continuous and

$$\mathbb{E}_{\theta_X,\theta_Y}(X|Y=y) = \sum_x x f_{\theta_X,\theta_Y}(x|y)dx$$

if X and Y are discrete.

Note that conditional expectations and conditional variances are RVs! This is because, e.g., the conditional expectation of $X|Y$ depends on Y and will vary randomly according to Y.

Here are some helpful properties of conditional expectations and variances, known as iterated expectations and variances:

$$\begin{aligned}
\mathbb{E}_{\theta_X} &= \mathbb{E}_{\theta_Y}\left(\mathbb{E}_{\theta_X,\theta_Y}(X|Y)\right) \\
\mathrm{Var}_{\theta_X} &= \mathbb{E}_{\theta_Y}\left(\mathrm{Var}_{\theta_X,\theta_Y}(X|Y)\right) + \mathrm{Var}_{\theta_Y}\left(\mathbb{E}_{\theta_X,\theta_Y}(X|Y)\right)
\end{aligned}$$

Here is a proof of iterated expectation, to better understand how/why "nesting" or "iterating" two expectations gives back one expectation. I will use the densities here, but a similar proof can be used for PMFs.

$$\begin{aligned}
\mathbb{E}_{\theta_Y}\left(\mathbb{E}_{\theta_X,\theta_Y}(X|Y)\right) &= \int \left(\int x f_{\theta_x,\theta_Y}(x|y)dx \right) f_{\theta_Y}(y)dy \text{ (using LOTUS)} \\
&= \int \int x f_{\theta_X,\theta_Y}(x,y)dxdy \\
&= \int x \left(\int f_{\theta_X,\theta_Y}(x,y)dy \right) dx \\
&= \int x f_{\theta_X} dx \\
&= \mathbb{E}_{\theta_X}(X).
\end{aligned}$$

Finally, let us briefly discuss multivariate distributions. Let $X = (X_1, X_2, \ldots, X_p)$ be a random vector (i.e., it is a vector where each entry is a RV). Similarly to the bivariate case, we can define the joint distribution, marginal distributions, and conditional distributions (where we can condition on one or more RVs). As you can imagine, it can get very complicated to work with multivariate distributions.

But suppose each X_j for $j = 1, 2, \ldots, p$ is independent and identically distributed with PDF/PMF $f_\theta(x_j)$. This is in general denoted by $X_1, X_2, \ldots, X_p \overset{\text{iid}}{\sim} f_\theta$. Then the problem is simplified:

$$f_\theta(x_1, x_2, \ldots, x_p) = \prod_{j=1}^{p} f_\theta(x_j).$$

2 Review of Introductory Statistics

The fundamental issue of statistics is that the previously defined probability distributions depend on some parameter θ, which is generally unknown. And we cannot actually understand an RV, the probability of events, etc. without knowing θ.

2.1 Sampling

We visualize that there is a population of values $x_1, x_2, \ldots, x_N$ in a population, which are non-random. For example, the population could be the heights of all Americans. The problem is that typically it is infeasible or even impossible to obtain the full set of values in the population.

As a result, we take a sample from the population. We aim for it to be random and representative of the population, to best capture the values in the population. If that is the case (and ignoring issues of sampling without replacement), then we denote it as $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta$, where n is the sample size. The goal is then to estimate θ or other parameters of interest (such as the mean, which is not necessarily equal to θ).

2.2 Estimation

If we use the random sample to estimate θ, then denote the resulting estimator as $\hat{\theta} = T(X_1, X_2, \ldots, X_n)$. Note that this estimator depends on the RVs in the random sample! Therefore, $\hat{\theta}$ is random too!

Due to the fact that we randomly sampled, $\hat{\theta}$ will not necessarily be perfectly representative of θ. Therefore, we as statisticians will clearly be concerned with the quality of the estimator.

There are a variety of ways that we can measure the quality of the estimator. Often, there are trade-offs between these properties; there is not necessarily an "ideal" estimator for all problems. So each property will be of varying levels of importance for each problem. We can look at both finite-sample properties, as well as asymptotic properties.

2.3 Finite-sample Properties of Estimators

If a property holds in "finite-samples", then we are saying that the property holds for any sample size n. Some of the most common finite-sample properties that we might care about are the following:

- size of the bias
- size of the standard error
- size of the MSE

Note that sometimes we wish to emphasize the dependence of the estimator on n, in which case we write it as $\hat{\theta}_n$, for estimating θ. We will now define the above terms.

Definition 2.1. Bias of an Estimator . The bias of the estimator $\hat{\theta}$ for θ is

$$bias(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta.$$

Thus, it measures how close the estimator is to the estimand on average. Intuitively, we would want the bias to be as small as possible (without compromising other properties, since as we will see soon, there can be trade-offs).

Definition 2.2. Unbiased Estimator. The estimator $\hat{\theta}$ is unbiased if $bias(\hat{\theta}) = 0$.

I will also emphasize again that if $\hat{\theta} = T_\theta(X_1, \ldots, X_n)$, then for it to be unbiased, we need its expectation to equal θ for ALL n. That is, we need $\mathbb{E}(\hat{\theta}_n) - \theta = 0$ for all n.

The variance of $\hat{\theta}$ uses the same variance definition from before, and as always the standard deviation is the square root of the variance. Since $\hat{\theta}$ is an estimator, we often refer to its standard deviation as its "standard error", denoted $se(\hat{\theta})$. Having a small variance is generally a desirable property, like having low bias. If we are considering using two different estimators, then the relative efficiency can be helpful to calculate.

Definition 2.3. Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two estimators for θ. Then their relative efficiency is

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

The MSE quantifies how much the estimator deviates (in terms of squared difference) from the estimand, on average.

Definition 2.4. The MSE is defined as

$$MSE(\hat{\theta}) = \mathbb{E}\left((\hat{\theta} - \theta)^2\right).$$

It can be shown that the MSE is the sum of the variance and the squared bias:

$$MSE(\hat{\theta}) = (bias(\hat{\theta}))^2 + \mathrm{Var}(\hat{\theta}).$$

A consequence of this fact is the "bias-variance tradeoff". Suppose the true MSE of an estimator is 200, for example. Then if the variance of the estimator is low, then that must mean the bias is high. On the other hand, if the bias of the estimator is low, then that must mean the variance is high. Where this tradeoff really becomes a big issue is in machine learning. This is not covered in this class, but basically, it is possible that the estimator from an algorithm is so flexible that it can "memorize" the specific dataset (getting extremely low bias) while changing significantly when applied to new data (getting extremely high variance). Of course, it is possible to just have a good estimator with low bias and low variance and thus low MSE, in which case there is not really a "tradeoff".

2.4 Types of Convergence

Some of the asymptotic properties of estimators rely on different types of convergence, so let us first define those.

Definition 2.5. Convergence in Probability. A sequence of RVs $W_1, W_2, \ldots, W_j, W_{j+1}, \ldots$ converges in probability to W (written $W_n \xrightarrow{p} W$) if: for every $\epsilon > 0$,

$$\mathbb{P}(|W_n - W| \geq \epsilon) = 0.$$

Here, $W = c$, a non-random constant.

For example, we can imagine randomly sampling Americans repeatedly and getting the sample mean of their heights. That is, our sequence would consist of the RVs $W_n = \frac{1}{n}\sum_{i=1}^n X_i$, where n is the sample size, for $n = 1, 2, \ldots$. Then we might consider whether W_n converges in probability to $W = \mu$, the population mean. (The benefit of using a sequence of RVs will become clearer when we study the asymptotic properties of estimators, i.e., the properties of estimators with increasing sample size n).

Definition 2.6. A sequence of random variables $X_1, X_2, \ldots$ converges in distribution to a RV X (denoted $X_n \implies X$) if

$$\lim_{n \to \infty} F_n(t) = F(t)$$

for all t for which F is continuous.

There are many other types of convergence too, such as almost sure convergence, convergence in mean square, etc., but this should suffice for now.

While we're here, let's review some important theorems related to convergence.

Imagine flipping a coin 5 times. Does it seem unlikely that we would get 100% heads? Not really – it was only 5 flips. What about if we flipped it 50 times? 500? 5000? Intuitively, we would expect that with more flips, we would be getting closer to 50% heads.

In fact, we have this theorem.

Theorem 2.1. Weak Law of Large Numbers (LLN). Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim}$ with mean $\mu = \mathbb{E}(X_i)$ and variance $\sigma^2 = \text{Var}(X_i)$. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then $\bar{X}_n$ converges in probability to μ as $n \to \infty$.

In the coin flip example, the proportion of flips in a particular sample is a mean (specifically, the mean for iid Bernoulli RVs), so the weak LLN applies. Therefore, it is true that with more flips, we get closer to the true proportion .5 of heads.

Possibly the most powerful theorems in (at least classical) statistics is the Central Limit Theorem (CLT).

Theorem 2.2. Central Limit Theorem (CLT). Let $X_1, X_2, \ldots \overset{\text{iid}}{\sim}$ with mean $\mu = \mathbb{E}(X_i)$ and variance $\text{Var}(X_i)$, and let $\bar{X}_n = \frac{1}{n}\bar{X}_n$. Then

$$\bar{X}_n \to N(\mu, \sigma^2/n).$$

We can equivalently write this as

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \to N(0,1).$$

We already know from the weak LLN that $\bar{X}_n$ will converge to μ. But the CLT goes two steps further, which makes it incredibly powerful! First, the normal distribution is well-known with many great properties; so as long as we have a mean or sum and a large enough sample size, we should be able to use a normal approximation. Additionally, the variance will shrink to 0 as the sample size increases! If you would like to learn more about the *why* behind the CLT, then the math channel 3Blue1Brown on YouTube has some fantastic videos on it: `https://www.youtube.com/watch?v=zeJD6dqJ5lo` and `https://www.youtube.com/watch?v=cy8r7WSuT1I`.

Lastly, let's look at two theorems that help us with the asymptotic behavior of *transformations* of RVs.

First, we can carry over asymptotic properties of a sequence of RVs to continuous functions of that RV, which can be quite helpful.

Theorem 2.3. Continuous Mapping Theorem. Let g be a continuous function. If W_n converges in probability to W, then $g(W_n)$ converges in probability to g(W). If X_n converges in distribution to X, then $g(X_n)$ converges in distribution to $g(X)$.

Secondly, we can figure out the asymptotics of two sequences of RVs under some assumptions.

Theorem 2.4. Slutsky's Theorem. Let $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be sequences of random variables. Suppose X_n converges in distribution to X and Y_n converges in probability to c, where c is some constant. Then

$$\begin{aligned} X_n + Y_n &\implies X + c \\ Y_n X_n &\implies cX \\ X_n / Y_n &\implies X/c. \end{aligned}$$

2.5 Asymptotic Properties of Estimators

We will discuss this next section.

3 Exercises

Problem 1.

Let $X_1, X_2, \dots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$, and let $\hat{\lambda}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Find the bias, standard error, and MSE.

Solution

Applying the definition of bias, we have

$$\begin{aligned} bias(\hat{\lambda}_n) &= \mathbb{E}(\hat{\lambda}_n) - \lambda \\ &= \lambda - \lambda = 0, \end{aligned}$$

since a sample mean is always unbiased for the population mean.

The variance is

$$\begin{aligned} \text{Var}(\hat{\lambda}_n) &= \frac{1}{n^2}\sum_{i=1}^n \text{Var}(X_i) \text{ (by a prop. of var.)} \\ &= \frac{1}{n^2} * n\lambda \text{ (var. of a Poisson RV is } \lambda) \\ &= \frac{\lambda}{n}. \end{aligned}$$

The MSE can be computed as

$$\begin{aligned} MSE(\hat{\lambda_n}) &= \left(bias(\hat{\lambda_n})\right)^2 + \text{Var}(\hat{\lambda}_n) \\ &= 0^2 + \frac{\lambda}{n} = \frac{\lambda}{n}. \end{aligned}$$

Problem 2.

Suppose $X_1, X_2 \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Show that $\bar{X}$ is normally distributed.

Solution

Probably the easiest way to show this is through MGFs:

$$\begin{aligned}
M_{\bar{X}}(t) &= \mathbb{E}(e^{t\bar{X}}) \text{ (by def.)} \\
&= \mathbb{E}(e^{t(X_1+X_2)/2} \\
&= \mathbb{E}(e^{tX_1/2}e^{tX_2/2}) \\
&= \mathbb{E}(e^{tX_1/2})(e^{tX_2/2}) \text{ (since } X_1, X_2 \text{ are ind., so functions of them are too)} \\
&= M_X(t/2)M_Y(t/2) \\
&= e^{\frac{\mu t}{2}} e^{\frac{\sigma^2 t}{2}\frac{1}{2}^2} e^{\frac{\mu t}{2}} e^{\frac{\sigma^2 t}{2}\frac{1}{2}^2} \\
&= e^{\mu} e^{\frac{\sigma^2 t^2}{2*2}},
\end{aligned}$$

which is the MGF of a normal RV with mean μ and variance $\sigma^2/2 = \sigma^2/n$. Thus, by uniqueness of MGFs, $\bar{X}$ is indeed normal.

Note that we can easily extend this proof to the case of n iid normal RVs.

It is actually also true that if we have $X_1, X_2, \ldots, X_n \sim N(\mu_i, \sigma_i^2)$ and scalars $a_1, a_2, \ldots, a_n$, then the linear combination $\sum_{i=1}^n a_i X_i$ is distributed as normal. But it is more difficult to prove.

Problem 3.

Let $X_1, \ldots, X_n$ be iid RVs with mean μ and variance σ^2 What does $\hat{\theta} = \frac{1}{\bar{X}}$ converge in probability to? Or is there not enough to say?

Solution

We know that by the weak LLN, $\bar{X}$ converges in probability to μ. Meanwhile, $\hat{\theta}$ is a function of $\bar{X}$: specifically, $\hat{\theta} = g(\bar{X}) = \frac{1}{\bar{X}}$. This function g is continuous as long as $\bar{X} \neq 0$. Assuming that is not the case, then by the Continuous Mapping Theorem, we have that $\hat{\theta}$ converges in probability to $\frac{1}{\mu}$.

Problem 4.

What is the intuition of convergence in probability, as well as convergence in distribution?

Solution

Convergence in probability: First of all, recall the definition of convergence of sequences from math.

Definition 3.1. Let $a_1, a_2, \ldots$ be a sequence. a_n converges to a value $a \in \mathbb{R}$ if for every $\delta > 0$, there exists a positive natural number N such that for all $n \geq N$,

$$|a_n - a| < \delta.$$

Intuitively, this means that no matter how small a value δ, we can find a point in the sequence such that the sequence values a_n deviate from a by no more than δ. That is, the sequence is getting arbitrarily close to a as it progresses! Other ways of writing that "a_n converges to a" are $\lim\limits_{n\to\infty} a_n = a$ or $a_n \to a$.

But we are in statistics, so dealing with RVs. That means the definition of convergence from math does not apply. We cannot ever say that an RV *exactly* equals something (unless the RV is trivially defined as a constant). We can only talk about the *probability* of the RV equalling certain values. Thus, we "wrap" the event that the "W_n deviates from W by more than ϵ" inside a probability $\mathbb{P}(\cdot)$:

$$\mathbb{P}(|W_n - W| \geq \epsilon).$$

So we now have a deterministic sequence and can talk about regular ol convergence from math! Specifically, we are saying that the probability converges to *zero*. That is, no matter how large a value $\epsilon > 0$, and no matter how small a value $\delta > 0$, we can always find a point in the sequence of W_n such that after that: W_n deviates from W by more than ϵ with probability less than δ. It is becoming increasingly unlikely that W_n deviates from W by more than ϵ as we proceed through the sequence.

Convergence in Distribution: Just like with convergence in probability, we need to think about how to translate statements about RVs to statements about deterministic values. As we know, the CDF of an RV is unique, so it would make sense to use that fact to define the notion of RVs getting close to another RV. Specifically, let $X_1, X_2 \ldots$ be a sequence of RVs, and X be an RV. Then we can define a sequence of CDFs $F_1, F_2, \ldots$, as well as F. And CDFs are deterministic, so we can apply definitions of convergence from math!

The only issue is that we can't exactly use the definition of convergence of sequences. That's because the CDFs are *functions*. However, recall that in math, there is a definition of convergence that applies to functions, which we can apply.

Definition 3.2. Let g_n be a sequence of real-valued functions, i.e., each $g_n : Y \to \mathbb{R}$ is a function where $Y \subset \mathbb{R}$. Let $g : Y \to \mathbb{R}$ be a function. Then g_n converges pointwise to g if: for every point $y \in Y$, we have that $g_n(y) \to g(y)$.

So basically, this definition has extended regular convergence of sequences to convergence of functions — by checking that for every point y in the functions' domain, the sequence $g_n(y)$ converges to $g(y)$.

Overall, convergence in distribution means that the sequence of RVs' CDFs converge pointwise to the CDF of another RV.

Problem 5.

This problem is an example of when Bayes' Theorem is useful. Suppose the probability of a rare cancer in the population is .03. A test for the cancer is known to have 88% accuracy if the patient has cancer and 93% accuracy if the patient does not have cancer. What is the probability that someone has cancer, given that their test if positive?

Solution

Let's define some quantities:

$$\begin{aligned} \mathbb{P}(+|C) &= .88 \\ \mathbb{P}(-|C^c) &= .93 \\ \mathbb{P}(C) &= .03 \\ P(C^c) &= .97. \end{aligned}$$

Applying Bayes' Theorem, we have

$$\mathbb{P}(C|+) = \frac{\mathbb{P}(+|C)\mathbb{P}(C)}{\mathbb{P}(+)}.$$

The only issue is that we don't know $\mathbb{P}(+)$. However, we can calculate it with the law of total probability:

$$\begin{aligned} \mathbb{P}(+) &= \mathbb{P}(+|C)\mathbb{P}(C) + \mathbb{P}(+|C^c)\mathbb{P}(C^c) && (2) \\ &= 0.88(.03) + 0.07(.97). && (3) \end{aligned}$$

Plugging everything in, we have

$$\mathbb{P}(C|+) = \frac{0.88(.03)}{0.88(.03) + 0.07(.97)} \approx 0.280.$$

So even though the test is 88% accurate for people with cancer and 93% accurate for people without cancer, if you get a positive result on the test, the probability of you actually having cancer is only about 0.28!

What's going on? Basically this is happening because the true incidence rate of cancer is so low. Rare events are very tricky to analyze statistically.

Later in the course, we will apply Bayes' Theorem extensively, when we cover Bayesian statistics.