MATH 157: Intermediate Probability

Connor Neely, Fall 2024

* Adapted from FA24 Lecture Notes.

1 Univariate Distributions

1.1 Discrete Random Variables

The fundamental object of this class is the random variable. We'll start with the discrete case.

Definition: Probability mass function

Let X be a discrete random variable. The function $P(X = k)$ is called a probability mass function if

- $P(X = k) \geq 0$ for all k and
- $\sum_k P(X = k) = 1$.

Before looking at some important kinds of probability mass functions, we'll examine some of their general characteristics regarding center and spread.

Definition: Expected value

The expected value of a discrete random variable X is given by

$$E(X) = \sum_k k\, P(X = k).$$

The expected value of X is also called the mean of X, and is sometimes denoted μ or μ_X. Also, for any function $g(x)$ we define

$$E[g(x)] = \sum_k g(k)\, P(X = k).$$

Note that the expected value may not exist if the above sums do not converge.

Definition: Variance and standard deviation

The variance of a random variable X with mean μ is the average squared distance from μ. That is,

$$\text{var}(X) = E[(x - \mu)^2].$$

It is typically denoted by σ^2 or σ_x^2. The standard deviation is

$$\text{sd}(x) = \sqrt{\text{var}(x)},$$

and is typically denoted by σ or σ_x.

There's a couple of theorems regarding expected value and variance which will prove useful in computations.

Theorem 1.1: Expected value and variance of linear expressions

For real a and b,

$$E(ax + b) = a\, E(x) + b \quad \text{and} \quad \text{var}(ax + b) = a^2\, \text{var}(x).$$

Theorem 1.2: Variance via expected value

For a random variable X,

$$\text{var}(X) = E(X^2) - E(X)^2.$$

We'll prove later that a random variable is almost always within two standard deviations of its mean, with probability at least 75%.

1.2 Some Important Discrete RVs

The simplest kind of random variable is a uniform random variable, in which all outcomes have equal probabilities. This is boring, though, so let's define a few more interesting ones.

Definition: Geometric random variable

Consider an infinite sequence of independent binary events, each of which has a probability p of success. If X is the index of the first success, then X is called a geometric random variable and

$$P(X = k) = (1-p)^{k-1}p.$$

We say that X obeys a geometric distribution with parameter p, or $X \sim \text{geo}(p)$.

Theorem 1.3: EV of a geometric random variable

A random variable $X \sim \text{geo}(p)$ has expected value $E(X) = 1/p$.

We omit a proof because the result is intuitive, but it involves differentiating the geometric series sum.

As a side note, we could compute $P(X \geq k)$ in a few ways: via a finite geometric series, by subtracting off $P(X < k)$, or by finding the probability $(1-p)^{k-1}$ that the first $k-1$ events are failures. Now onto another!

Definition: Binomial random variable

Consider a sequence of n independent binary events, each of which has a probability p of success. If X is the number of successes in this sequence, then X is called a binomial random variable and

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

We say that X obeys a geometric distribution with parameters n and p, or $X \sim \text{bin}(n, p)$.

Theorem 1.4: EV and variance of a binomial random variable

A random variable $X \sim \text{bin}(n, p)$ has $E(X) = np$ and $\text{var}(X) = np(1-p)$.

Another intuitive result, and we'll defer a proof to later. For our last variable, we start with $X \sim \text{bin}(n, p)$ and take the large and small limits of n and p, respectively. This gives us the following.

Definition: Poisson random variable

Consider a long sequence of rare binary events with expected value $\lambda > 0$. If X is the number of successes in the sequence, then X is called a Poisson random variable and

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

We say that X obeys a Poisson distribution with parameter λ, or $X \sim \text{poi}(\lambda)$.

Theorem 1.5: Variance of a Poisson random variable

A random variable $X \sim \text{poi}(\lambda)$ has $\text{var}(X) = \lambda$.

The full proof of this is lengthy. As a jumping-off point, it involves computing $E(X^2) = E[X(X-1) + X]$.

1.3 Conditional Probability

We'll finish off our discrete discussion by looking at conditional probability.

Definition: Conditional probability

For two events A and B, the probability of A given B is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

From this definition we can write a few intuitive identities.

Theorem 1.6: Conditional probability identities

For two events A and B, the following are true.

(a) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

(b) $P(A \cap B) = P(A)P(B)$ for independent A, B.

(c) $P(A) = P(A \cap B) + P(A \cap B^c)$.

Theorem 1.7: Law of total probability

For any A and disjoint $B_1, \cdots, B_n$,

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A \mid B_i)P(B_i).$$

Everything here is pretty straightforward, but now we'll look at how it all ties into more complicated scenarios.

Example: Craps

Suppose we roll two dice to get a total X. There are three outcomes: if $X = 7, 11$ then we win immediately; if $X = 2, 3, 12$ we lose immediately; and if X is anything else we continue rolling until X appears again, in which case we win, or until 7 appears, in which case we lose. The probability of winning is

$$P(\text{win}) = \sum_{k=2}^{12} P(\text{win} \mid X = k)\, P(X = k).$$

This is easy enough for the first two cases, so we'll focus on the third. In particular, we can compute $P(\text{win} \mid X = 4)$ in three different ways. Let Π_4 denote the probability of rolling another 4 before a 7.

- We could simply use a geometric series:

$$\Pi_4 = \frac{3}{36} + \left(\frac{27}{36}\right)\left(\frac{3}{36}\right) + \left(\frac{27}{36}\right)^2\left(\frac{3}{36}\right) + \cdots = \frac{1}{3}.$$

- We could also solve the equation

$$\Pi_4 = \frac{3}{36} + \frac{27}{36}\Pi_4 \implies \Pi_4 = \frac{1}{3}.$$

- Finally, we could simply observe that there are three ways to roll a 4 and six ways to roll a 7. Thus $\Pi_4 = 3/(3+6) = 1/3$.

Doing the same for all of the other possibilities will yield $P(\text{win}) = 0.49\overline{29}$.

Craps is very close to being a fair game, but for repeated bets the odds are deceptive.

Theorem 1.8: Gambler's ruin

Consider a game with probability p of success. A player making repeated \$1 bets starts with \$$i$ and aims to reach \$$n$, with $0 \leq i \leq n$. The probability of reaching this goal is given by

$$a_i = \frac{1-(q/p)^i}{1-(q/p)^n},$$

where $p \neq 1/2$ and $q = 1-p$.

Proving this involves solving the recurrence

$$a_i = pa_{i+1} + qa_{i-1}, \quad a_0 = 0,\ a_n = 1;$$

For a fair game ($p = 1/2$) we get $a_i = i/n$, and in any other case with $p \neq 0$ we get the result above.

1.4 Continuous Random Variables

Now we'll make our discussion a bit more interesting by considering the continuous case.

Definition: Probability density function

A continuous random variable is described by a probability density function f_X (or just f or $f(x)$) such that, for any $a \leq b$,

$$P(a \leq X \leq b) = \int_a^b f_X(x)\,dx.$$

Such a function must satisfy $f_X(x) \geq 0$ for all x and $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$.

The key characteristics we have for discrete random variables generalize nicely to continuous ones. Note that the expected value may not exist if the distribution has "heavy tails", in which case the variance also doesn't exist.

Definition: Expected value

For a continuous random variable X,

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx, \qquad E[g(x)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx.$$

The variance and standard deviation are the same as we defined before: $\mathrm{var}(X) = E[(x-\mu)^2]$, where μ is the distribution's mean. Also, all of the nice properties we found about these characteristics hold, too! Namely:

$$E(ax+b) = a\,E(x) + b, \qquad \mathrm{var}(ax+b) = a^2\mathrm{var}(x), \qquad \mathrm{var}(X) = E(X^2) - E(X)^2.$$

By the linearity of integration we could generalize the first result to arbitrary linear combinations of functions. Now we'll make one more definition that we really could've made earlier in the discrete case but also wouldn't have been useful until now.

Definition: Cumulative distribution function

For any random variable X, we define the cumulative distribution function

$$F_X(x) = P(X \leq x).$$

Note that for discrete X this is defined as a summation, while for continuous X it's an integral.

These functions are useful not only in their own right, but also in working with functions of a random variable: given a pdf for X and an expression for Y in terms of X, we can easily generate a pdf for Y. This is best illustrated via an example.

Example: A function of a random variable

Consider a continuous random variable with pdf $f_X(x)$ defined on $0 \leq x \leq 30$, and let $Y = \frac{9}{5}X + 32$. To write down a pdf for Y we first compute its feasible region:

$$0 \leq x \leq 30 \implies 32 \leq \frac{9}{5}x + 32 \leq 86 \implies 32 \leq y \leq 86.$$

Now we look to the cumulative distribution functions, writing $F_Y(y)$ in terms of F_X:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P\left(\frac{9}{5}X + 32 \leq y\right) \\ &= P\left(X \leq \frac{5}{9}(y - 32)\right) \\ &= F_X\left(\frac{9}{5}(y - 32)\right). \end{aligned}$$

Finally, we can differentiate both sides with respect to y to get the pdf

$$f_Y(y) = \frac{9}{5} f_X\left(\frac{9}{5}(y - 32)\right).$$

1.5 Some Important Continuous RVs

Now we'll look at some particular examples of continuous random variables. Like before, we can start simple with the uniform random variable.

Definition: Uniform random variable

X is a uniform random variable with parameters $a < b$ if it has a pdf

$$f(x) = \frac{1}{b - a}.$$

Such a random variable is denoted $X \sim u(a, b)$.

Theorem 1.9: EV and variance of a uniform RV

If $X \sim u(a, b)$ then

$$E(X) = \frac{a + b}{2}, \qquad \text{var}(X) = \frac{(b - a)^2}{12}.$$

What's interesting about this is that we might treat an arbitrary distribution function $F_X(x)$ as a probability distribution, note that it ranges from 0 to 1, and find that $F_X(x) \sim u(0, 1)$. We might restate this in a slightly different way to make it more intuitive.

Theorem 1.10

If $U \sim u(0, 1)$, then $y = F_X^{-1}(U)$ has the same pdf as X. (It follows that a number U generated from $u(0, 1)$ represents the $100u$-th percentile of X.)

Next, we have the continuous analog of the geometric distribution.

Definition: Exponential random variable

X is an exponential random variable with parameter $\lambda > 0$ if X has pdf

$$f(x) = \lambda e^{-\lambda x}.$$

Such a random variable is denoted $X \sim \text{expo}(\lambda)$.

λ is often called the rate parameter, and $1/\lambda$ represents the "average time" between rare events. In this way, the λ here is the same as in the Poisson distribution!

Theorem 1.11: EV and variance of an exponential RV

If $X \sim \text{expo}(\lambda)$, then

$$E(X) = \frac{1}{\lambda}.$$

Also, $E(X^2) = 2/\lambda^2$, so

$$\text{var}(X) = E(X^2) - E(X)^2 = \frac{1}{\lambda^2}.$$

Finally, we have what is perhaps the most important random variable of all. Given its significance, we'll take a slightly closer look at its key properties.

Definition: Normal random variable

X is a normal random variable with parameters μ and σ^2 if it has pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty.$$

Such a random variable is denoted $X \sim n(\mu, \sigma^2)$.

This is a complicated pdf! To simplify it slightly, we could use the change of variables $z = (x - \mu)/\sigma$ to get

$$f_Z(z) = f_X(\sigma z + \mu) \cdot \sigma = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

This is the pdf of the standard normal, $Z \sim n(0, 1)$; we call it $\phi(z)$. It still has no elementary antiderivative, though, so in practice we map problems to this standard normal's cdf and use numerical methods to calculate any probabilities. (It's useful to bear in mind that approximately 68% of the distribution is contained within σ of the mean, approximately 95% within 2σ, and 99.7% within 3σ.)

Theorem 1.12: Transforming normal RVs

If $X \sim n(\mu, \sigma^2)$ and $Y = aX + b$, then

(a) $E(Y) = a\mu + b$ and

(b) $\text{var}(Y) = a^2\sigma^2$.

In fact, Y is also normal with these parameters.

Later we'll also prove that linear combinations of independent normal RVs X_i are, in a way, "preserved"; in particular,

$$\sum_i (a_i X_i + b) \sim n\left(\sum_i (a_i\mu_i + b), \sum_i a_i^2\sigma_i^2\right).$$

Importantly, if we have two identical but independent RVs X_1, X_2, something like $X_1 + X_2$ does not have the same distribution as $2X_1$—their variances are different!

2 Multivariate Distributions

2.1 Joint Probability Functions

So far we've been talking about distributions of individual variables. Now we'll look at the slightly more general case of two variables.

Definition: Joint probability mass function

Let X and Y be discrete random variables. The function $P(X = x, Y = y)$ is a joint PMF if

- $P(X = x,\ Y = y) \geq 0$ for all x, y and
- $\sum_{x,y} P(X = x, Y = y) = 1$.

The marginal PMF for X is

$$P(X = x) = \sum_{y} P(X = x, Y = y);$$

X and Y are independent if $P(X = x, Y = y) = P(X = x)P(Y = y)$ for all x, y.

Definition: Joint probability density function

Let X and Y be continuous random variables. The function $f_{X,Y}(x, y)$ is a joint PDF if

- $f_{X,Y}(x, y) \geq 0$ for all x, y and
- $\iint_{\mathbb{R}^2} f_{X,Y}(x, y)\, dxdy = 1$.

the marginal PDF for X is

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x, y)\, dy;$$

X and Y are independent if $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for all x, y.

Also, note that knowing the marginals of two random variables is not necessarily enough to obtain their full joint PF. To do this, we must also require that X and Y are independent.

Definition: Conditional probability distribution

Let X and Y be random variables. For discrete and continuous variables we have, respectively,

$$P(Y = y \,|\, X = x) = \frac{P(Y = y,\ X = x)}{P(X = x)}, \qquad f_{Y|X}(y \,|\, x) = \frac{f_{X,Y}(x, y)}{f_X(x)}.$$

2.2 Multivariate Distributions

We can generalize the above discussion to probability functions of several random variables. Rather than give a general treatment for the discrete case, though, we'll focus on one particularly important distribution.

Definition: Multinomial distribution

If an experiment with m outcomes (with probabilities $p_1, \ldots, p_m$) is independently performed n times and X_i is the number of times outcome i occurs, then

$$P(X_1 = x_1, \ldots, X_m = x_m) = \binom{n}{x_1, \ldots, x_m} p_1^{x_1} \cdots p_m^{x_m}, \qquad \binom{n}{x_1, \ldots, x_m} = \frac{n!}{x_1! \cdots x_m!},$$

where the coefficient is called a multinomial coefficient.

We'll spend much more time looking at the continuous case.

Definition: Continuous multivariate distribution

If $X_1, \ldots, X_n$ have joint pdf $f_{X_1,\ldots,X_n}$ then

$$P\big((x_1, \ldots, x_n) \in A\big) = \int_A f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\, dx_1 \cdots dx_n.$$

The marginal density of X_1 is found by "integrating out" $X_2, \ldots, X_n$. $X_1, \ldots, X_n$ are independent if

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

with a rectangular feasible region.

When working with univariate distributions we were often tasked with transforming from one variable to another. We'll see the same problems with multivariate distributions, and the procedure generalizes nicely. When transforming from several variables to one variable we simply write the cdf for the new variable and differentiate to get its pdf—the methods for doing this might be more involved, but the general idea is there.

We can do the same when we transform to several variables. In the single-variable case, if $Y = g(X)$ and g is invertible then

$$f_Y(y) = f_X\big(g^{-1}(y)\big) \left| \frac{d}{dy} g^{-1}(y) \right| = f_X(x) \left| \frac{dx}{dy} \right|.$$

More generally, say we want to transform from some $\mathbf{X}$ to some $\mathbf{Y}$ given a $\mathbf{y} = \mathbf{g}(\mathbf{x})$. We start by inverting $\mathbf{g}$ and writing $\mathbf{x} = \mathbf{h}(\mathbf{y})$, use this to rewrite the pdf of $\mathbf{X}$ in terms of $\mathbf{h}(\mathbf{y})$, and compute

$$\mathbf{f_Y}(\mathbf{y}) = \mathbf{f_X}\big(\mathbf{h}(\mathbf{y})\big)\, |J|.$$

Here $J = \det(d\mathbf{h}/d\mathbf{y})$ is the Jacobian determinant of the transformation from $\mathbf{Y}$ to $\mathbf{X}$:

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix}.$$

Finally, don't forget the feasible region!

So far we've seen a very natural generalization of the two-variable case discussed earlier. We can get similarly natural results about expected value.

Definition: Expected value

If $\mathbf{X} = (X_1, \ldots, X_n)$ have joint PDF $f_{\mathbf{X}}$ and $g : \mathbb{R}^n \to \mathbb{R}$ is a real-valued function, we define

$$E[g(\mathbf{x})] = \int_{\mathbb{R}^n} g(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})\, dx_1 \cdots dx_n.$$

Theorem 2.1: Expected value of a sum

For any random variables $X_1, \ldots, X_n$,

$$E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n).$$

Note that this holds when $X_1, \ldots, X_n$ are not independent. If they are independent, then we get something familiar.

Theorem 2.2: Expected value of a product

If $X_1, \ldots, X_n$ are independent, then

$$E(X_1 \cdots X_n) = E(X_1) \cdots E(X_n).$$

We can leverage these results to solve some neat problems.

Example: The matching problem

Suppose n students do an assignment and their instructor wants them to grade each other's work. We'd like to find the expected number of students who get their own submission back, assuming random distribution (under the condition that each student grades one submission).

We can define an "indicator variable"

$$X_i = \begin{cases} 1 & \text{student } i \text{ gets own submission,} \\ 0 & \text{otherwise,} \end{cases}$$

so $E(X_i) = 1/n$. We might think of the assignment distribution process as a sequence of X_is. The X_i are not independent, but we can still say that

$$E(X) = E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n) = 1,$$

where X is the number of students who get their own submission.

We could follow a very similar line of reasoning to prove what our intuition previously told us about the binomial expected value.

2.3 Variance and Covariance

Now we seek similarly convenient results about variance. To make this happen, we'll need some scaffolding.

Definition: Covariance

The covariance of X and Y is

$$\operatorname{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)],$$

where $E(X) = \mu_X$ and $E(Y) = \mu_Y$.

Theorem 2.3: Covariance via expected value

For any random variables X and Y,

$$\operatorname{cov}(X, Y) = E(XY) - E(X)E(Y).$$

It follows that the covariance of two independent variables is zero, and that the covariance of a variable with itself (a scenario with maximum dependence) is simply its variance.

This suggests some intuition as to what the covariance is quantifying. If $\text{cov}(X,Y) > 0$ then we say X, Y are positively correlated, meaning higher values of X are likely to correspond to higher Y. If $\text{cov}(X,Y) < 0$ then we analogously have negative correlation, and if $\text{cov}(X,Y) = 0$ then X, Y are uncorrelated.

We'll come back to this in a moment. First, let's look at some of the nice results covariance gives us.

Theorem 2.4: Covariance is bilinear

For any random variables X, Y, Z,

(a) $\text{cov}(aX, bY) = ab\,\text{cov}(X,Y)$ and

(b) $\text{cov}(X, Y+Z) = \text{cov}(X,Y) + \text{cov}(X,Z)$.

Theorem 2.5: Variance of a sum

For any random variables $X_1, \ldots, X_n$,

$$\text{var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} \text{cov}(X_i, X_j)$$

Corollary 2.6

For independent random variables $X_1, \ldots, X_n$,

$$\text{var}(X_1 + \cdots + X_n) = \text{var}(X_1) + \cdots + \text{var}(X_n).$$

At this point we could use indicator variables to show that, for the matching problem, $\text{var}(X) = 1$. We could similarly derive the variance of a binomial random variable. More novel, though, is a connection we can make to correlation.

Definition: Correlation

Let X, Y have variances σ_X^2, σ_Y^2. The correlation between X and Y is

$$\rho(x,y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}.$$

Theorem 2.7: Restrictions on correlations

For any random variables X, Y,

(a) $|\rho(X,Y)| \leq 1$ and

(b) $\rho(X,Y) = \pm 1$ if and only if Y is a linear function of X.

Proof. Variance is strictly non-negative, so we have

$$\begin{aligned} 0 &\leq \text{var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) \\ &= \frac{1}{\sigma_X^2}\text{var}(X) + \frac{1}{\sigma_Y^2}\text{var}(Y) + \frac{2\text{cov}(X,Y)}{\sigma_X \sigma_Y} \\ &= 1 + 1 + 2\rho(X,Y), \end{aligned}$$

meaning $\rho(X,Y) \geq -1$. If we instead started with $\text{var}(X/\sigma_X - Y/\sigma_Y)$ we'd get $\rho(X,Y) \leq 1$. This proves statement (a). If we'd used equalities rather than inequalities we'd get the forward direction for (b); the reverse is just some algebra. □

2.4 Moment Generating Functions

We'll look at one last really important characteristic of random variables. We state uniqueness without proof.

Definition: Moment generating function

A random variable X has moment generating function

$$M_X(t) = E[e^{tX}].$$

Theorem 2.8: Uniqueness of an MGF

If $M_{X_1}(t) = M_{X_2}(t)$ then $f_{X_1}(x) = f_{X_2}(x)$.

Notice, in general, that

$$M_X^{(n)}(t) = E[X^n e^{tX}] \quad \text{and} \quad M_X^{(n)}(0) = E[X^n]$$

for $n \geq 0$. We will sometimes refer to $E[X^n]$ as the nth moment of X.

It'll be important for later to know that the MGF corresponding to $X \sim n(0,1)$ is $M_X(t) = e^{t^2/2}$. We could use this next result to quickly obtain the MGF for a general normal distribution.

Theorem 2.9

If $Y = aX + b$ then $M_Y(t) = e^{bt} M_X(at)$.

Finally, a way to find the MGF of a sum of independent variables.

Theorem 2.10: MGF of a sum

If $X_1, \ldots, X_n$ are independent, where X_i has MGF $M_{X_i}(t)$, then $Y = X_1 + \cdots + X_n$ has MGF

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t).$$

We could use this result with indicator variables to determine the MGF of, say, the binomial distribution. We could also use it to prove the claim we made earlier about summing independent normal random variables.

2.5 Theoretical Results

To finish things off, we'll look at some of the more important theoretical results that are accessible to us now.

Theorem 2.11: Markov's inequality

For any non-negative random variable X, and for any real $t > 0$,

$$P(X \geq t) \leq \frac{E(X)}{t}.$$

Proof. Fix $t > 0$ and write, for a discrete random variable,

$$E(x) = \sum_{k \geq 0} k\, P(X = k) \geq \sum_{k \geq t} k\, P(X = k) \geq \sum_{k \geq t} t\, P(X = k) = t\, P(X \geq t).$$

So $P(X \geq t) \leq E(X)/t$, as desired. The continuous case is completely analogous. □

Theorem 2.12: Chebyshev's inequality

If $E(X) = \mu$, then for any $t > 0$,

$$P(|X - \mu| \geq t) \leq \frac{\text{var}(X)}{t^2}.$$

Proof. Let $Y = (X - \mu)^2$. Note that $Y \geq 0$ and $E(Y) = \text{var}(X)$, so using Markov's inequality we can write

$$P(|x - \mu| \geq t) = P(Y \geq t^2) \leq \frac{E(Y)}{t^2} = \frac{\text{var}(X)}{t^2},$$

as desired. □

We were able to get this slightly more powerful result from a less powerful one because of our added assumption that the variance exists. Chebyshev's inequality is a workhorse in probability theory, but we can also prove something much more familiar.

Definition: Sample mean

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables. Their sample mean is

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}.$$

Theorem 2.13: EV and variance of a sample mean

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables, each with mean μ and variance σ^2. Then

$$E(\overline{X}_n) = \mu, \qquad \text{var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

Theorem 2.14: Law of large numbers

$\overline{X}_n$ converges to μ in probability ($\overline{X}_n \xrightarrow{P} \mu$). That is, for any $\epsilon > 0$,

$$\lim_{n \to \infty} P\Big(|\overline{X}_n - \mu| < \varepsilon\Big) = 1.$$

Proof. For $\epsilon > 0$, by Chebyshev's inequality

$$P\Big(|\overline{X}_n - \mu| \geq \epsilon\Big) \leq \frac{\text{var}(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{N\epsilon^2}.$$

and

$$P\Big(|\overline{X}_n - \mu| < \epsilon\Big) = 1 - \frac{\sigma^2}{N\epsilon^2}.$$

Thus

$$\lim_{n \to \infty} P\Big(|\overline{X}_n - \mu| < \varepsilon\Big) = \lim_{n \to \infty} \left(1 - \frac{\sigma^2}{N\epsilon^2}\right) = 1,$$

as desired. □

2.6 The Central Limit Theorem

For our last theoretical result, we have one of the foundational theorems of probability and statistics.

Theorem 2.15: Central limit theorem

Suppose the random variables $X_1, \ldots, X_n$ are independent and identically distributed with mean μ and variance σ^2. The sample mean $\overline{X}_n$ is normally distributed with mean μ and variance σ^2/n.

Proof. Define

$$Y = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}.$$

We will show that $M_Y(t) \to e^{t^2/2}$ as $n \to \infty$ so that Y is a standard normal random variable. To this end, let $Y_i = (X_i - \mu)/\sigma$ so that

$$\sum_{i=1}^{n} \frac{Y_i}{\sqrt{n}} = \frac{\sum X_i - n\mu}{\sigma\sqrt{n}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = Y.$$

Now we need only find the moment generating function of each Y_i. We can easily show that $E(Y_i) = 0$ and $\text{var}(Y_i) = 1$, meaning $E(Y_i^2) = 1$ and by Taylor's theorem the MGF of Y_i is

$$\begin{aligned} M(t) &= M(0) + M'(0)t + M''(0)\frac{t^2}{2} + M'''(c)\frac{t^3}{6} \\ &= 1 + E(Y_i)t + E(Y_i^2)\frac{t^2}{2} + M'''(c)\frac{t^3}{6} \\ &= 1 + \frac{t^2}{2} + M'''(c)\frac{t^3}{6} \end{aligned}$$

for some $0 < c < t$. Now note that a random variable X/a has MGF $M_X(t/a)$ for any constant a, so we can write

$$\begin{aligned} M\left(\frac{t}{\sqrt{n}}\right) &= 1 + \frac{(t/\sqrt{n})^2}{2} + \frac{M'''(c_n)}{6}(t/\sqrt{n})^3 \\ &= 1 + \frac{t^2}{2n} + \frac{d_n t^3}{n\sqrt{n}} \\ &= 1 + \frac{t^2/2 + d_n t^3/\sqrt{n}}{n}, \end{aligned}$$

where we'de defined $d_n = M'''(c_n)/6$. We could show that d_n is bounded, meaning the numerator above tends to $t^2/2$ as $n \to \infty$. Thus

$$M_Y(t) = \left(1 + \frac{t^2/2}{n}\right)^n = e^{t^2/2},$$

as desired. □