1. Distributions

Dave Goldsman

H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

3/2/20

Georgia Tech

Outline

Lesson 1.1 — Bernoulli and Binomial Distributions

Goal: We'll discuss lots of interesting distributions in this module.

The module will be a compendium of results, some of which we'll prove, and some of which we've already seen previously.

Special emphasis will be placed on the **Normal distribution**, because it's so important and has so many implications, including the **Central Limit Theorem**.

In the next few lessons we'll discuss some important **discrete** distributions:

- Bernoulli and Binomial Distributions
- Hypergeometric Distribution
- Geometric and Negative Binomial Distributions
- Poisson Distribution

Definition: The **Bernoulli distribution** with parameter p is given by

$$X = \begin{cases} 1 & \text{w.p. } p \text{ ("success")} \\ 0 & \text{w.p. } q \text{ ("failure")} \end{cases}$$

Recall: $\mathrm{E}[X] = p$, $\mathrm{Var}(X) = pq$, and $M_X(t) = pe^t + q$.

Definition: The **Binomial distribution** with parameters n and p is given by

$$P(Y = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, \ldots, n.$$

Georgia Tech

Example: Toss 2 dice and take the sum; repeat 5 times. Let Y be the number of 7's you see. $Y \sim \text{Bin}(5, 1/6)$. Then, e.g.,

$$P(Y = 4) \; = \; \binom{5}{4}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^{5-4}. \quad \Box$$

Theorem: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(p) \Rightarrow Y \equiv \sum_{i=1}^n X_i \sim \text{Bin}(n, p)$.

Proof: This kind of result can easily be proved by a moment generating function uniqueness argument, as we mentioned in previous modules. $\Box$

Think of the Binomial as the number of successes from n Bern(p) trials.

Theorem: $Y \sim \text{Bin}(n, p)$ implies

$$\mathrm{E}[Y] \;=\; \mathrm{E}\Big[\sum_{i=1}^{n} X_i\Big] \;=\; \sum_{i=1}^{n} \mathrm{E}[X_i] \;=\; np.$$

Similarly,

$$\mathrm{Var}(Y) \;=\; npq.$$

We've already seen that

$$M_Y(t) \;=\; (pe^t + q)^n.$$

Theorem: Certain Binomials add up: If $Y_1, \ldots, Y_k$ are *independent* and $Y_i \sim \text{Bin}(n_i, p)$, then

$$\sum_{i=1}^{k} Y_i \;\sim\; \text{Bin}\Big(\sum_{i=1}^{k} n_i, p\Big).$$

Georgia Tech

Lesson 1.2 — Hypergeometric Distribution

Definition: You have a objects of type 1 and b objects of type 2.

Select n objects *without replacement* from the $a+b$.

Let X be the number of type 1's selected. Then X has the **Hypergeometric distribution** with pmf

$$P(X=k) \;=\; \frac{\binom{a}{k}\binom{b}{n-k}}{\binom{a+b}{n}}, \quad k = 0, 1, \ldots, \min(a, n-b, n).$$

Example: 25 sox in a box. 15 red, 10 blue. Pick 7 w/o replacement.

$$P(\text{exactly 3 reds are picked}) \;=\; \frac{\binom{15}{3}\binom{10}{4}}{\binom{25}{7}}. \quad \square$$

Theorem: After some algebra, it turns out that

$$\mathrm{E}[X] \;=\; n\left(\frac{a}{a+b}\right) \quad \text{and}$$

$$\mathrm{Var}(X) \;=\; n\left(\frac{a}{a+b}\right)\left(1-\frac{a}{a+b}\right)\left(\frac{a+b-n}{a+b-1}\right).$$

Remark: Here, $\frac{a}{a+b}$ here plays the role of p in the Binomial distribution.

And then the corresponding $Y \sim \mathrm{Bin}(n,p)$ results would be

$$\mathrm{E}[Y] \;=\; n\left(\frac{a}{a+b}\right) \quad \text{and}$$

$$\mathrm{Var}(Y) \;=\; n\left(\frac{a}{a+b}\right)\left(1-\frac{a}{a+b}\right).$$

So same mean as the Hypergeometric, but slightly larger variance.

Lesson 1.3 — Geometric and Negative Binomial Distributions

Definition: Suppose we consider an infinite sequence of independent Bern(p) trials.

Let Z equal the number of trials *until the first success* is obtained. The event $Z = k$ corresponds to $k - 1$ failures, and then a success. Thus,

$$P(Z = k) \;=\; q^{k-1}p, \quad k = 1, 2, \ldots,$$

and we say that Z has the **Geometric distribution** with parameter p.

Notation: $X \sim \text{Geom}(p)$.

We'll get the mean and variance of the Geometric via the mgf....

Theorem: The mgf of the Geom(p) is

$$M_Z(t) = \frac{pe^t}{1-qe^t}, \quad \text{for } t < \ell\text{n}(1/q).$$

Proof:

$$\begin{aligned} M_Z(t) &= \mathrm{E}[e^{tZ}] \\ &= \sum_{k=1}^{\infty} e^{tk} q^{k-1} p \\ &= pe^t \sum_{k=0}^{\infty} (qe^t)^k \\ &= \frac{pe^t}{1-qe^t}, \quad \text{for } qe^t < 1. \quad \Box \end{aligned}$$

Corollary: $\mathrm{E}[Z] = 1/p$.

Proof:

$$\begin{aligned}
\mathrm{E}[Z] &= \frac{d}{dt} M_Z(t)\Big|_{t=0} \\
&= \frac{(1-qe^t)(pe^t) - (-qe^t)(pe^t)}{(1-qe^t)^2}\Big|_{t=0} \\
&= \frac{pe^t}{(1-qe^t)^2}\Big|_{t=0} \\
&= \frac{p}{(1-q)^2} = \frac{1}{p}. \quad \Box
\end{aligned}$$

Remark: We could also have proven this directly from the definition of expected value.

Similarly, after a lot of algebra,

$$\mathrm{E}[Z^2] \;=\; \frac{d^2}{dt^2} M_Z(t)\Big|_{t=0} \;=\; \frac{2-p}{p^2},$$

and then

$$\mathrm{Var}(Z) \;=\; \mathrm{E}[Z^2] - (\mathrm{E}[Z])^2 \;=\; q/p^2. \quad \Box$$

Example: Toss a die repeatedly. What's the probability that we observe a 3 for the first time on the 8th toss?

Answer: The number of tosses we need is $Z \sim \mathrm{Geom}(1/6)$.
$P(Z = 8) = q^{k-1}p = (5/6)^7(1/6)$. $\Box$

How many tosses would we expect to take?

Answer: $\mathrm{E}[Z] = 1/p = 6$ tosses. $\Box$

Memoryless Property of Geometric

Theorem: Suppose $Z \sim \mathrm{Geom}(p)$. Then for positive integers s, t, we have

$$P(Z > s + t | Z > s) \;=\; P(Z > t).$$

Why is it called the Memoryless Property? If an event hasn't occurred by time s, the probability that it will occur after an additional t time units is the same as the (unconditional) probability that it will occur after time t — it forgot that it made it past time s!

Georgia Tech

Proof: First of all, for any $t = 0, 1, 2, \ldots$, the tail probability is

$$P(Z > t) \;=\; P(t \text{ Bern}(p) \text{ failures in a row}) \;=\; q^t. \qquad (*)$$

Then

$$\begin{aligned} P(Z > s + t | Z > s) &= \frac{P(Z > s + t \cap Z > s)}{P(Z > s)} \\ &= \frac{P(Z > s + t)}{P(Z > s)} \\ &= \frac{q^{s+t}}{q^s} \quad \text{(by } (*)\text{)} \\ &= q^t \\ &= P(Z > t). \quad \square \end{aligned}$$

Example: Let's toss a die until a 5 appears for the first time. Suppose that we've already made 4 tosses without success. What's the probability that we'll need more than 2 additional tosses before we observe a 5?

Let Z be the number of tosses required. By the Memoryless Property (with $s = 4$ and $t = 2$) and $(*)$, we want

$$P(Z > 6 | Z > 4) \; = \; P(Z > 2) \; = \; (5/6)^2. \quad \square$$

Fun Fact: The Geom(p) is the only discrete distribution with the memoryless property.

Not-as-Fun Fact: Some books define the Geom(p) as the number of Bern(p) *failures* until you observe a success. $\#$ failures $= \#$ trials $- 1$. You should be aware of this inconsistency, but don't worry about it now.

Definition: Suppose we consider an infinite sequence of independent Bern(p) trials.

Now let W equal the number of trials *until the rth success* is obtained. $W = r, r+1, \ldots$. The event $W = k$ corresponds to exactly $r-1$ successes by time $k-1$, and then the rth success at time k.

We say that W has the **Negative Binomial distribution** (aka the Pascal distribution) with parameters r and p.

Example: 'FFFFSFS' corresponds to $W = 7$ trials until the $r = 2$nd success.

Notation: $W \sim \text{NegBin}(r, p)$.

Remark: As with the Geom(p), the exact definition of the NegBin depends on what book you're reading.

Theorem: If $Z_1, \ldots, Z_r \overset{\text{iid}}{\sim} \text{Geom}(p)$, then $W = \sum_{i=1}^r Z_i \sim \text{NegBin}(r,p)$. In other words, Geometrics add up to a NegBin.

Proof: Won't do it here, but you can use the mgf technique. $\Box$

Anyhow, it makes sense if you think of Z_i as the number of trials after the $(i-1)$st success up to and including the ith success.

Since the Z_i's are i.i.d., the above theorem gives:

$$\begin{aligned} \text{E}[W] &= r\text{E}[Z_i] = r/p, \\ \text{Var}(W) &= r\text{Var}(Z_i) = rq/p^2, \\ M_W(t) &= [M_{Z_i}(t)]^r = \left(\frac{pe^t}{1-qe^t}\right)^r. \end{aligned}$$

Just to be complete, let's get the pmf of W. Note that $W = k$ iff you get exactly $r-1$ successes by time $k-1$, and then the rth success at time k. So...

$$P(W = k) = \left[\binom{k-1}{r-1}p^{r-1}q^{k-r}\right]p = \binom{k-1}{r-1}p^r q^{k-r}, \quad k = r, r+1, \ldots$$

Example: Toss a die until a 5 appears for the third time. What's the probability that we'll need exactly 7 tosses?

Let W be the number of tosses required. Clearly, $W \sim \mathrm{NegBin}(3, 1/6)$.

$$P(W = 7) = \binom{7-1}{3-1}(1/6)^3(5/6)^{7-3}.$$

How are the Binomial and NegBin Related?

$X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} \text{Bern}(p) \;\Rightarrow\; Y \equiv \sum_{i=1}^{n} X_i \sim \text{Bin}(n, p).$

$Z_1, \ldots, Z_r \stackrel{\text{iid}}{\sim} \text{Geom}(p) \;\Rightarrow\; W \equiv \sum_{i=1}^{r} Z_i \sim \text{NegBin}(r, p).$

$\text{E}[Y] = np$, $\text{Var}(Y) = npq$.

$\text{E}[W] = r/p$, $\text{Var}(W) = rq/p^2$.

Lesson 1.4 — Poisson Distribution

We'll first talk about **Poisson processes**.

Let $N(t)$ be a **counting process**. That is, $N(t)$ is the number of occurrences (or arrivals, or events) of some process over the time interval $[0, t]$. $N(t)$ looks like a step function.

Examples: $N(t)$ could be any of the following.

(a) Cars entering a shopping center (by time t).
(b) Defects on a wire (of length t).
(c) Raisins in cookie dough (of volume t).

Let $\lambda > 0$ be the average number of occurrences per unit time (or length or volume).

In the above examples, we might have:
(a) $\lambda = 10$/min. (b) $\lambda = 0.5$/ft. (c) $\lambda = 4/\text{in}^3$.

A Poisson process is a specific counting process....

First, some notation: $o(h)$ is a generic function that goes to zero faster than h goes to zero.

Definition: A **Poisson process (PP)** is one that satisfies the following:

(i) There is a short enough interval of time h, such that, for all t,

$$\begin{aligned} P(N(t+h) - N(t) = 0) &= 1 - \lambda h + o(h) \\ P(N(t+h) - N(t) = 1) &= \lambda h + o(h) \\ P(N(t+h) - N(t) \geq 2) &= o(h) \end{aligned}$$

(ii) The distribution of the **"increment"** $N(t+h) - N(t)$ only depends on the length h.

(iii) If $a < b < c < d$, then the two "increments" $N(d) - N(c)$ and $N(b) - N(a)$ are *independent* RVs.

English translation of Poisson process assumptions.

(i) Arrivals basically occur one-at-a-time, and then at rate λ/unit time. (We must make sure that λ doesn't change over time.)

(ii) The arrival pattern is *stationary* — it doesn't change over time.

(iii) The numbers of arrivals in two disjoint time intervals are independent

Poisson Process Example: Neutrinos hit a detector. Occurrences are rare enough so that they really do happen one-at-a-time. You never get arrivals of groups of neutrinos. Further, the rate doesn't vary over time, and all arrivals are independent of each other. $\square$

Anti-Example: Customers arrive at a restaurant. They show up in groups, not one-at-a-time. The rate varies over the day (more at dinnertime). Arrivals may not be independent. This ain't a Poisson process. $\square$

Definition: Let X be the number of occurrences in a Poisson(λ) process in a *unit interval* of time. Then X has the **Poisson distribution** with parameter λ.

Notation: $X \sim \text{Pois}(\lambda)$.

Theorem/Definition: $X \sim \text{Pois}(\lambda) \Rightarrow P(X = k) = e^{-\lambda}\lambda^k/k!$, $k = 0, 1, 2, \ldots$.

Proof: The proof follows from the PP assumptions and involves some simple differential equations.

To begin with, let's define $P_x(t) \equiv P(N(t) = x)$, i.e., the probability of exactly x arrivals by time t.

Note that the probability that there haven't been any arrivals by time $t + h$ can be written in terms of the probability that there haven't been any arrivals by time t. . . .

$$
\begin{aligned}
&P_0(t+h) \\
&\quad = \; P(N(t+h)=0) \\
&\quad = \; P(\text{no arrivals by time } t \text{ and then no arrivals by time } t+h) \\
&\quad = \; P\big(\{N(t)=0\} \cap \{N(t+h)-N(t)=0\}\big) \\
&\quad = \; P(N(t)=0)P(N(t+h)-N(t)=0) \quad \text{(by indep increments (iii))} \\
&\quad \doteq \; P(N(t)=0)(1-\lambda h) \quad \text{(by (i) and a little bit of (ii))} \\
&\quad = \; P_0(t)(1-\lambda h).
\end{aligned}
$$

Thus,

$$\frac{P_0(t+h)-P_0(t)}{h} \;\doteq\; -\lambda P_0(t).$$

Taking the limit as $h \to 0$, we have

$$P_0'(t) \;=\; -\lambda P_0(t). \tag{1}$$

Similarly, for $x > 0$, we have

$$\begin{aligned}
P_x(t+h) &= P(N(t+h) = x) \\
&= P(N(t+h) = x \text{ and no arrivals during } [t, t+h]) \\
&\quad + P(N(t+h) = x \text{ and } \geq 1 \text{ arrival during } [t, t+h]) \\
&\quad \text{(Law of Total Probability)} \\
&\doteq P\big(\{N(t) = x\} \cap \{N(t+h) - N(t) = 0\}\big) \\
&\quad + P\big(\{N(t) = x - 1\} \cap \{N(t+h) - N(t) = 1\}\big) \\
&\quad \text{(by (i), only consider case of one arrival in } [t, t+h]) \\
&= P(N(t) = x)P(N(t+h) - N(t) = 0) \\
&\quad + P(N(t) = x - 1)P(N(t+h) - N(t) = 1) \\
&\quad \text{(by independent increments (iii))} \\
&\doteq P_x(t)(1 - \lambda h) + P_{x-1}(t)\lambda h.
\end{aligned}$$

Taking the limits as $h \to 0$, we obtain

$$P'_x(t) \;=\; -\lambda\big[P_{x-1}(t) - P_x(t)\big], \quad x = 1, 2, \ldots. \tag{2}$$

The solution of differential equations (1) and (2) is easily shown to be

$$P_x(t) \;=\; \frac{(\lambda t)^x e^{-\lambda t}}{x!}, \quad x = 0, 1, 2, \ldots.$$

(If you don't believe me, just plug in and see for yourself!)

Noting that $t = 1$ for the Pois(λ) finally completes the proof. $\square$

Remark: λ can be changed simply by changing the units of time.

Examples:

X = # calls to a switchboard in 1 minute $\sim$ Pois(3 / min)

Y = # calls to a switchboard in 5 minutes $\sim$ Pois(15 / 5 min)

Z = # calls to a switchboard in 10 sec $\sim$ Pois(0.5 / 10 sec)

Theorem: $X \sim \text{Pois}(\lambda) \Rightarrow$ mgf is $M_X(t) = e^{\lambda(e^t-1)}$.

Proof:

$$M_X(t) \;=\; \mathrm{E}[e^{tX}] \;=\; \sum_{k=0}^{\infty} e^{tk}\left(\frac{e^{-\lambda}\lambda^k}{k!}\right) \;=\; e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^t)^k}{k!} \;=\; e^{-\lambda}e^{\lambda e^t}. \quad \square$$

Theorem: $X \sim \text{Pois}(\lambda) \Rightarrow \mathrm{E}[X] = \text{Var}(X) = \lambda$.

Proof (using mgf):

$$\mathrm{E}[X] \;=\; \frac{d}{dt}M_X(t)\Big|_{t=0} \;=\; \frac{d}{dt}e^{\lambda(e^t-1)}\Big|_{t=0} \;=\; \lambda e^t M_X(t)\Big|_{t=0} \;=\; \lambda.$$

Similarly,

$$\begin{aligned}
\mathrm{E}[X^2] &= \frac{d^2}{dt^2} M_X(t)\Big|_{t=0} \\
&= \frac{d}{dt}\left(\frac{d}{dt} M_X(t)\right)\Big|_{t=0} \\
&= \lambda \frac{d}{dt}\left(e^t M_X(t)\right)\Big|_{t=0} \\
&= \lambda\left[e^t M_X(t) + e^t \frac{d}{dt} M_X(t)\right]\Big|_{t=0} \\
&= \lambda e^t\left[M_X(t) + \lambda e^t M_X(t)\right]\Big|_{t=0} \\
&= \lambda(1+\lambda).
\end{aligned}$$

Thus, $\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = \lambda(1+\lambda) - \lambda^2 = \lambda.$ $\square$

Georgia Tech

Example: Calls to a switchboard arrive as a Poisson process with rate 3 calls/min.

Let X = number of calls in 1 minute.

So $X \sim \text{Pois}(3)$, $\text{E}[X] = \text{Var}(X) = 3$, $P(X \leq 4) = \sum_{k=0}^{4} e^{-3}3^k/k!$.

Let Y = number of calls in 40 sec.

So $Y \sim \text{Pois}(2)$, $\text{E}[Y] = \text{Var}(Y) = 2$, $P(Y \leq 4) = \sum_{k=0}^{4} e^{-2}2^k/k!$. $\square$

Georgia Tech

Theorem (Additive Property of Poissons): Suppose $X_1, \ldots, X_n$ are *independent* with $X_i \sim \text{Pois}(\lambda_i)$, $i = 1, \ldots, n$. Then

$$Y \equiv \sum_{i=1}^{n} X_i \sim \text{Pois}\Big(\sum_{i=1}^{n} \lambda_i\Big).$$

Proof: Since the X_i's are independent, we have

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} e^{\lambda_i(e^t-1)} = e^{(\sum_{i=1}^{n} \lambda_i)(e^t-1)},$$

which is the mgf of the $\text{Pois}\big(\sum_{i=1}^{n} \lambda_i\big)$ distribution. $\square$

Example: Cars driven by males [females] arrive at a parking lot according to a Poisson process with a rate of $\lambda_1 = 3\,/\,\text{hr}$ [$\lambda_2 = 5\,/\,\text{hr}$]. All arrivals are independent.

What's the probability of exactly 2 arrivals in the next 30 minutes?

The total number of arrivals is $\text{Pois}(\lambda_1 + \lambda_2 = 8/\text{hr})$, and so the total in the next 30 minutes is $X \sim \text{Pois}(4)$. So $P(X = 2) = e^{-4}4^2/2!$. $\Box$

Lesson 1.5 — Uniform, Exponential, and Friends

Definition: The RV X has the **Uniform distribution** if it has pdf and cdf

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad F(x) = \begin{cases} 0, & x < a \\ \frac{x}{b-a}, & a \le x \le b \\ 1, & x \ge b. \end{cases}$$

Notation: $X \sim \text{Unif}(a, b)$.

Previous work showed that

$$\text{E}[X] = \frac{a+b}{2} \quad \text{and} \quad \text{Var}(X) = \frac{(a-b)^2}{12}.$$

We can also derive the mgf,

$$M_X(t) = \text{E}[e^{tX}] = \int_a^b e^{tx} \frac{1}{b-a}\, dx = \frac{e^{tb} - e^{ta}}{t(b-a)}.$$

Definition: The **Exponential(λ) distribution** has pdf

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Previous work showed that the cdf $F(x) = 1 - e^{-\lambda x}$,

$$\mathrm{E}[X] = 1/\lambda, \quad \text{and} \quad \mathrm{Var}(X) = 1/\lambda^2.$$

We also derived the mgf,

$$M_X(t) = \mathrm{E}[e^{tX}] = \int_0^\infty e^{tx} f(x)\,dx = \frac{\lambda}{\lambda - t}, \quad t < \lambda.$$

Memoryless Property of Exponential

Theorem: Suppose that $X \sim \text{Exp}(\lambda)$. Then for positive s, t, we have

$$P(X > s + t | X > s) \;=\; P(X > t).$$

Similar to the discrete Geometric distribution, the probability that X will survive an additional t time units is the (unconditional) probability that it will survive at least t — it forgot that it made it past time s! It's always "like new"!

Proof:

$$\begin{aligned} P(X > s + t | X > s) &= \frac{P(X > s + t \cap X > s)}{P(X > s)} = \frac{P(X > s + t)}{P(X > s)} \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X > t). \quad \square \end{aligned}$$

Example: Suppose that the life of a lightbulb is exponential with a mean of 10 months. If the light survives 20 months, what's the probability that it'll survive another 10?

$$P(X > 30 | X > 20) \;=\; P(X > 10) \;=\; e^{-\lambda x} \;=\; e^{-(1/10)(10)} \;=\; e^{-1}. \quad \square$$

Example: If the time to the next bus is exponentially distributed with a mean of 10 minutes, and you've already been waiting 20 minutes, you can expect to wait 10 more. $\square$

Remark: The exponential is the *only* continuous distribution with the Memoryless Property.

Remark: Look at $\mathrm{E}[X]$ and $\mathrm{Var}(X)$ for the Geometric distribution and see how they're similar to those for the exponential. (Not a coincidence.)

Definition: If X is a cts RV with pdf $f(x)$ and cdf $F(x)$, then its **failure rate function** is

$$S(t) \equiv \frac{f(t)}{P(X > t)} = \frac{f(t)}{1 - F(t)},$$

which can loosely be regarded as X's instantaneous rate of death, given that it has so far survived to time t.

Example: If $X \sim \text{Exp}(\lambda)$, then $S(t) = \lambda e^{-\lambda t}/e^{-\lambda t} = \lambda$. So if X is the exponential lifetime of a lightbulb, then its instantaneous burn-out rate is always λ — always good as new! This is clearly a result of the Memoryless Property. $\Box$

The Exponential is also related to the Poisson!

Theorem: Let X be the amount of time until the first arrival in a Poisson process with rate λ. Then $X \sim \text{Exp}(\lambda)$.

Proof:

$$\begin{aligned} F(x) &= P(X \le x) \;=\; 1 - P(\text{no arrivals in } [0,x]) \\ &= 1 - \frac{e^{-\lambda x}(\lambda x)^0}{0!} \quad \text{(since \# arrivals in } [0,x] \text{ is Pois}(\lambda x)) \\ &= 1 - e^{-\lambda x}. \quad \square \end{aligned}$$

Theorem: Amazingly, it can be shown (after a lot of work) that the interarrival times of a PP are *all* iid $\text{Exp}(\lambda)$! See for yourself when you take a stochastic processes course.

Example: Suppose that arrivals to a shopping center are from a PP with rate $\lambda = 20$/hr. What's the probability that the time between the 13th and 14th customers will be at least 4 minutes?

Let the time between customers 13 and 14 be X. Since we have a PP, the interarrivals are iid Exp(λ = 20/hr), so

$$P(X > 4\text{ min}) \;=\; P(X > 1/15\text{ hr}) \;=\; e^{-\lambda t} \;=\; e^{-20/15}. \quad \Box$$

Definition/Theorem: Suppose $X_1, \ldots, X_k \stackrel{\text{iid}}{\sim} \text{Exp}(\lambda)$, and let $S = \sum_{i=1}^k X_i$. Then S has the **Erlang distribution** with parameters k and λ, denoted $S \sim \text{Erlang}_k(\lambda)$.

The Erlang is simply the sum of iid exponentials.

Special Case: $\text{Erlang}_1(\lambda) \sim \text{Exp}(\lambda)$.

The pdf and cdf of the Erlang are

$$f(s) = \frac{\lambda^k e^{-\lambda s} s^{k-1}}{(k-1)!}, \quad s \geq 0, \quad \text{and} \quad F(s) = 1 - \sum_{i=0}^{k-1} \frac{e^{-\lambda s}(\lambda s)^i}{i!}.$$

Notice that the cdf is the sum of a bunch of Poisson probabilities. (Won't do it here, but this observation helps in the derivation of the cdf.)

Expected value, variance, and mgf:

$$\begin{aligned} \mathrm{E}[S] &= \mathrm{E}\Big[\sum_{i=1}^{k} X_i\Big] = \sum_{i=1}^{k} \mathrm{E}[X_i] = k/\lambda \\ \mathrm{Var}(S) &= k/\lambda^2 \\ M_S(t) &= \left(\frac{\lambda}{\lambda - t}\right)^k. \end{aligned}$$

Example: Suppose X and Y are iid Exp(2). Find $P(X+Y<1)$.

$$P(X+Y<1) = 1 - \sum_{i=0}^{k-1} \frac{e^{-\lambda s}(\lambda s)^i}{i!} = 1 - \sum_{i=0}^{2-1} \frac{e^{-(2\cdot 1)}(2\cdot 1)^i}{i!} = 0.594. \quad \square$$

Definition: X has the **Gamma distribution** with parameters $\alpha > 0$ and $\lambda > 0$ if it has pdf

$$f(x) \ = \ \frac{\lambda^\alpha e^{-\lambda x} x^{\alpha-1}}{\Gamma(\alpha)}, \ \ x \geq 0,$$

where

$$\Gamma(\alpha) \ \equiv \ \int_0^\infty t^{\alpha-1} e^{-t} \, dt$$

is the *gamma function*.

Remark: The Gamma distribution generalizes the Erlang distribution (where α has to be a positive integer). It has the same expected value and variance as the Erlang, with α in place of k.

Remark: If α is a positive integer, then $\Gamma(\alpha) = (\alpha - 1)!$. Party trick: $\Gamma(1/2) = \sqrt{\pi}$.

Lesson 1.6 — Other Continuous Distributions

Triangular(a, b, c) Distribution — good for modeling RVs on the basis of limited data (minimum, mode, maximum).

$$f(x) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)}, & a < x \le b \\ \frac{2(c-x)}{(c-b)(c-a)}, & b < x < c \\ 0, & \text{otherwise.} \end{cases}$$

$$\mathrm{E}[X] = \frac{a+b+c}{3} \quad \text{and} \quad \mathrm{Var}(X) = \text{mess}$$

Beta(a, b) Distribution — good for modeling RVs that are restricted to an interval.

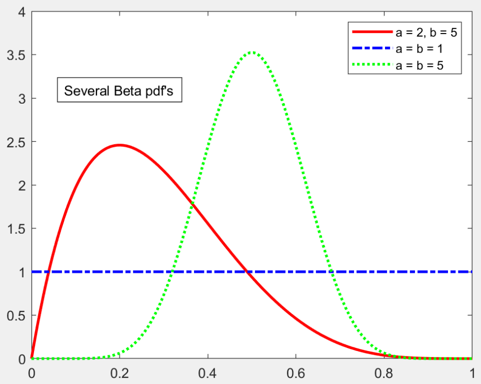
$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, x^{a-1}(1-x)^{b-1}, \ 0 < x < 1.$$

$$\mathrm{E}[X] = \frac{a}{a+b} \quad \text{and} \quad \mathrm{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}.$$

This distribution gets its name from the *beta function*, which is defined as

$$\beta(a, b) \equiv \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 x^{a-1}(1-x)^{b-1}\, dx.$$

The Beta distribution is very flexible. Here are a few family portraits.



Remark: Certain versions of the Uniform ($a = b = 1$) and Triangular distributions are special cases of the Beta.

Weibull(a, b) Distribution — good for modeling reliability models. a is the "scale" parameter, and b is the "shape" parameter.

$$\begin{aligned} f(x) &= ab(ax)^{b-1}e^{-(ax)^b}, \quad x > 0. \\ F(x) &= 1 - \exp[-(ax)^b], \quad x > 0. \\ \mathrm{E}[X] &= (1/a)\Gamma(1 + (1/b)) \quad \text{and} \quad \mathrm{Var}(X) \ = \ \text{slight mess} \end{aligned}$$

Remark: The Exponential is a special case of the Weibull.

Example: Time-to-failure T for a transmitter has a Weibull distribution with parameters $a = 1/(200 \text{ hrs})$ and $b = 1/3$. Then

$$\mathrm{E}[T] \ = \ 200\Gamma(1 + 3) \ = \ 1200 \text{ hrs}.$$

The probability that it fails before 2000 hrs is

$$F(2000) \ = \ 1 - \exp[-(2000/200)^{1/3}] \ = \ 0.884. \quad \Box$$

Cauchy Distribution — A "fat-tailed" distribution good for disproving things!

$$f(x) = \frac{1}{\pi(1+x^2)} \quad \text{and} \quad F(x) = \frac{1}{2} + \frac{\arctan(x)}{\pi}, \quad x \in \mathbb{R}.$$

Theorem: The Cauchy distribution has an undefined mean and infinite variance!

Weird Fact: $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ Cauchy $\Rightarrow \sum_{i=1}^n X_i/n \sim$ Cauchy. Even if you take the average of a bunch of Cauchys, you're right back where you started!

Alphabet Soup of Other Distributions

χ^2 distribution — coming up when we talk Statistics

t distribution — coming up

F distribution — coming up

Pareto, LaPlace, Rayleigh, Gumbel, Johnson distributions

Etc. . . .

Lesson 1.7 — Normal Distribution: Basics

The Normal Distribution is so important that we're giving it an entire section.

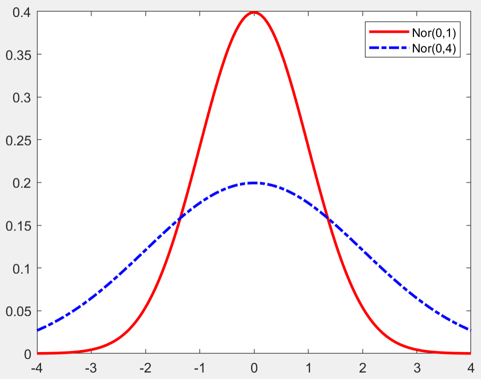
Definition: $X \sim \text{Nor}(\mu, \sigma^2)$ if it has pdf

$$f(x) \ = \ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right], \quad \forall x \in \mathbb{R}.$$

Remark: The Normal distribution is also called the Gaussian distribution.

Examples: Heights, weights, SAT scores, crop yields, and averages of things tend to be normal.

The pdf $f(x)$ is "bell-shaped" and symmetric around $x = \mu$, with tails falling off quickly as you move away from μ.



Small σ^2 corresponds to a "tall, skinny" bell curve; large σ^2 gives a "short, fat" bell curve.

Fun Fact (1): $\int_{\mathbb{R}} f(x)\,dx = 1$.

Proof: Transform to polar coordinates. Good luck. $\square$

Fun Fact (2): The cdf is

$$F(x) \;=\; \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(t-\mu)^2}{2\sigma^2}\right] dt \;=\; ??$$

Remark: No closed-form solution for this. Stay tuned.

Fun Facts (3) and (4): $\mathrm{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

Proof: Integration by parts or mgf (below).

Fun Fact (5): If X is *any* Normal RV, then

$$\begin{aligned} P(\mu - \sigma < X < \mu + \sigma) &= 0.6827 \\ P(\mu - 2\sigma < X < \mu + 2\sigma) &= 0.9545 \\ P(\mu - 3\sigma < X < \mu + 3\sigma) &= 0.9973. \end{aligned}$$

So almost all of the probability is contained within 3 standard deviations of the mean. (This is sort of what Toyota is referring to when it brags about "six-sigma" quality.)

Fun Fact (6): The mgf is $M_X(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$.

Proof: Calculus (or look it up in a table of integrals).

Theorem (Additive Property of Normals): If $X_1, \ldots, X_n$ are *independent* with $X_i \sim \mathrm{Nor}(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$, then

$$Y \equiv \sum_{i=1}^{n} a_i X_i + b \sim \mathrm{Nor}\Big(\sum_{i=1}^{n} a_i \mu_i + b, \sum_{i=1}^{n} a_i^2 \sigma_i^2\Big).$$

So a linear combination of independent normals is itself normal.

Proof: Since Y is a linear function,

$$
\begin{aligned}
M_Y(t) &= M_{\sum_i a_i X_i + b}(t) \;=\; e^{tb} M_{\sum_i a_i X_i}(t) \\
&= e^{tb} \prod_{i=1}^{n} M_{a_i X_i}(t) \quad (X_i\text{'s independent}) \\
&= e^{tb} \prod_{i=1}^{n} M_{X_i}(a_i t) \quad (\text{mgf of linear function}) \\
&= e^{tb} \prod_{i=1}^{n} \exp\left[\mu_i (a_i t) + \frac{1}{2}\sigma_i^2 (a_i t)^2\right] \quad (\text{normal mgf}) \\
&= \exp\left[\Big(\sum_{i=1}^{n} \mu_i a_i + b\Big) t + \frac{1}{2}\Big(\sum_{i=1}^{n} a_i^2 \sigma_i^2\Big) t^2\right],
\end{aligned}
$$

and we are done by mgf uniqueness. $\square$

Remark: A normal distribution is *completely characterized* by its mean and variance.

By the above, we know that a linear combination of independent normals is still normal. Therefore, when we add up independent normals, all we have to do is figure out the mean and variance — the normality of the sum comes for free.

Example: $X \sim \text{Nor}(3, 4)$, $Y \sim \text{Nor}(4, 6)$, and X, Y are independent. Find the distribution of $2X - 3Y$.

Solution: This is *normal* with

$$\text{E}[2X - 3Y] \;=\; 2\text{E}[X] - 3\text{E}[Y] \;=\; 2(3) - 3(4) \;=\; -6$$

and

$$\text{Var}(2X - 3Y) \;=\; 4\text{Var}(X) + 9\text{Var}(Y) \;=\; 70.$$

Thus, $2X - 3Y \sim \text{Nor}(-6, 70)$. $\square$

Corollary (of Additive Property Theorem):

$$X \sim \text{Nor}(\mu, \sigma^2) \;\Rightarrow\; aX + b \sim \text{Nor}(a\mu + b, a^2\sigma^2).$$

Proof: Immediate from the Additive Property after noting that $\text{E}[aX + b] = a\mu + b$ and $\text{Var}(aX + b) = a^2\sigma^2$. $\square$

Corollary (of Corollary):

$$X \sim \text{Nor}(\mu, \sigma^2) \;\Rightarrow\; Z \equiv \frac{X - \mu}{\sigma} \sim \text{Nor}(0, 1).$$

Proof: Use above Corollary with $a = 1/\sigma$ and $b = -\mu/\sigma$. $\square$

The manipulation described in this corollary is referred to as **standardization**.

Lesson 1.8 — Standard Normal Distribution

Definition: The $\text{Nor}(0, 1)$ is called the **standard normal distribution**, and is often denoted by Z.

The $\text{Nor}(0, 1)$ is nice because there are tables available for its cdf.

You can standardize any normal RV X into a standard normal by applying the transformation $Z = (X - \mu)/\sigma$. Then you can use the cdf tables.

The pdf of the $\text{Nor}(0, 1)$ is

$$\phi(z) \equiv \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad z \in \mathbb{R}.$$

The cdf is

$$\Phi(z) \equiv \int_{-\infty}^{z} \phi(t)\, dt, \quad z \in \mathbb{R}.$$

Remarks: The following results are easy to derive, usually via symmetry arguments.

$$
\begin{aligned}
P(Z \le a) &= \Phi(a) \\
P(Z \ge b) &= 1 - \Phi(b) \\
P(a \le Z \le b) &= \Phi(b) - \Phi(a)
\end{aligned}
$$

$$
\begin{aligned}
\Phi(0) &= 1/2 \\
\Phi(-b) &= P(Z \le -b) = P(Z \ge b) = 1 - \Phi(b) \\
P(-b \le Z \le b) &= \Phi(b) - \Phi(-b) = 2\Phi(b) - 1
\end{aligned}
$$

Then

$$P(\mu - k\sigma \le X \le \mu + k\sigma) = P(-k \le Z \le k) = 2\Phi(k) - 1.$$

So the probability that *any* normal RV is within k standard deviations of its mean doesn't depend on the mean or variance.

Famous $\text{Nor}(0, 1)$ table values. You can memorize these. Or you can use software calls, like `NORMDIST` in Excel (which calculates the cdf for *any* normal distribution.)

z	$\Phi(z) = P(Z \leq z)$
0.00	0.5000
1.00	0.8413
1.28	$0.8997 \approx 0.90$
1.645	0.9500
1.96	0.9750
2.33	$0.9901 \approx 0.99$
3.00	0.9987
4.00	≈ 1.0000

By the earlier "Fun Facts" and then the discussion on the last two pages, the probability that *any* normal RV is within k standard deviations of its mean is

$$P(\mu - k\sigma \le X \le \mu + k\sigma) \;=\; 2\Phi(k) - 1.$$

For $k = 1$, this probability is $2(0.8413) - 1 = 0.6827$.

There is a 95% chance that a normal observation will be within 2 s.d.'s of its mean.

99.7% of all observations are within 3 standard deviations of the mean!

Finally, note that Toyota's six-sigma corresponds to $2\Phi(6) - 1 \doteq 1.0000$.

Famous **Inverse** $\text{Nor}(0,1)$ table values. Can also use software, such as Excel's `NORMINV` function, which actually calculates inverses for *any* normal distribution, not just standard normal.

$\Phi^{-1}(p)$ is the value of z such that $\Phi(z) = p$. $\Phi^{-1}(p)$ is called the **pth quantile** of Z.

p	$\Phi^{-1}(p)$
0.90	1.28
0.95	1.645
0.975	1.96
0.99	2.33
0.995	2.58

Example: $X \sim \text{Nor}(21, 4)$. Find $P(19 < X < 22.5)$. **Standardizing,** we get

$$\begin{aligned} &P(19 < X < 22.5) \\ &\quad = \quad P\left(\frac{19-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{22.5-\mu}{\sigma}\right) \\ &\quad = \quad P\left(\frac{19-21}{2} < Z < \frac{22.5-21}{2}\right) \\ &\quad = \quad P(-1 < Z < 0.75) \\ &\quad = \quad \Phi(0.75) - \Phi(-1) \\ &\quad = \quad \Phi(0.75) - [1 - \Phi(1)] \\ &\quad = \quad 0.7734 - [1 - 0.8413] \;=\; 0.6147. \qquad \Box \end{aligned}$$

Example: Suppose that heights of men are $M \sim \text{Nor}(68, 4)$ and heights of women are $W \sim \text{Nor}(65, 1)$.

Select a man and woman *independently* at random.

Find the probability that the woman is taller than the man.

Answer: Note that

$$\begin{aligned} W - M &\sim \text{Nor}(\text{E}[W-M], \text{Var}(W-M)) \\ &\sim \text{Nor}(65-68, 1+4) \sim \text{Nor}(-3, 5). \end{aligned}$$

Then

$$\begin{aligned} P(W > M) &= P(W - M > 0) \\ &= P\left(Z > \frac{0+3}{\sqrt{5}}\right) \\ &= 1 - \Phi(3/\sqrt{5}) \\ &\doteq 1 - 0.910 = 0.090. \quad \Box \end{aligned}$$

Lesson 1.9 — Sample Mean of Normals

The **sample mean** of $X_1, \ldots, X_n$ is $\bar{X} \equiv \sum_{i=1}^{n} X_i/n$.

Corollary (of old theorem):
$X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2) \Rightarrow \bar{X} \sim \text{Nor}(\mu, \sigma^2/n)$.

Proof: By previous work, as long as $X_1, \ldots, X_n$ are iid something, we have $\text{E}[\bar{X}] = \mu$ and $\text{Var}(\bar{X}) = \sigma^2/n$. Since $\bar{X}$ is a linear combination of independent normals, it's also normal. Done. $\square$

Remark: This result is *very significant*! As the number of observations increases, $\text{Var}(\bar{X})$ gets *smaller* (while $\text{E}[\bar{X}]$ remains constant). In fact, it's called the **Law of Large Numbers**.

In the upcoming statistics portion of the course, we'll learn that this makes $\bar{X}$ an excellent **estimator** for the mean μ, which is typically unknown in practice.

Georgia Tech

Example: Suppose that $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} \text{Nor}(\mu, 16)$. Find the sample size n such that

$$P(|\bar{X} - \mu| \le 1) \;\ge\; 0.95.$$

How many observations should you take so that $\bar{X}$ will have a good chance of being close to μ?

Solution: Note that $\bar{X} \sim \text{Nor}(\mu, 16/n)$. Then

$$\begin{aligned} P(|\bar{X} - \mu| \le 1) &= P(-1 \le \bar{X} - \mu \le 1) \\ &= P\left(\frac{-1}{4/\sqrt{n}} \le \frac{\bar{X} - \mu}{4/\sqrt{n}} \le \frac{1}{4/\sqrt{n}}\right) \\ &= P\left(\frac{-\sqrt{n}}{4} \le Z \le \frac{\sqrt{n}}{4}\right) \\ &= 2\Phi(\sqrt{n}/4) - 1. \end{aligned}$$

Now we have to find n such that this probability is at least 0.95....

$2\Phi(\sqrt{n}/4) - 1 \geq 0.95$ iff

$\Phi(\sqrt{n}/4) \geq 0.975$ iff

$$\frac{\sqrt{n}}{4} \;\geq\; \Phi^{-1}(0.975) \;=\; 1.96$$

iff $n \geq 61.47$ or 62.

So if you take the average of 62 observations, then $\bar{X}$ has a 95% chance of being within 1 of μ. $\square$

Lesson 1.10 — The Central Limit Theorem + Proof

The Central Limit Theorem is the most-important theorem in probability and statistics!

CLT: Suppose $X_1, \ldots, X_n$ are iid with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. Then as $n \to \infty$,

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - \mathrm{E}[\bar{X}]}{\sqrt{\mathrm{Var}(\bar{X})}} \xrightarrow{d} \mathrm{Nor}(0,1),$$

where " $\xrightarrow{d}$ " means that the cdf of $Z_n \to$ the $\mathrm{Nor}(0,1)$ cdf.

Slightly Honors Informal Proof: Suppose that the mgf $M_X(t)$ of the X_i's exists and satisfies certain technical conditions that you don't need to know about. (OK, $M_X(t)$ has to exist around $t = 0$, among other things.)

Moreover, without loss of generality (since we're standardizing anyway) and for notational convenience, we'll assume that $\mu = 0$ and $\sigma^2 = 1$.

We will be done if we can show that the mgf of Z_n converges to the mgf of $Z \sim \text{Nor}(0,1)$, i.e., we need to show that $M_{Z_n}(t) \to e^{t^2/2}$ as $n \to \infty$.

To get things going, the mgf of Z_n is

$$\begin{aligned} M_{Z_n}(t) &= M_{\sum_{i=1}^n X_i/\sqrt{n}}(t) \\ &= M_{\sum_{i=1}^n X_i}(t/\sqrt{n}) \quad \text{(mgf of a linear function of a RV)} \\ &= \left[M_X(t/\sqrt{n})\right]^n \quad (X_i\text{'s are iid}). \end{aligned}$$

Georgia Tech

Thus, taking logs, our goal is to show that

$$\lim_{n\to\infty} n\,\ell\mathrm{n}\big(M_X(t/\sqrt{n})\big) \;=\; t^2/2.$$

If we let $y \;=\; 1/\sqrt{n}$, our revised goal is to show that

$$\lim_{y\to 0} \frac{\ell\mathrm{n}\big(M_X(ty)\big)}{y^2} \;=\; t^2/2.$$

Before proceeding further, note that

$$\lim_{y\to 0} \ell\mathrm{n}\big(M_X(ty)\big) \;=\; \ell\mathrm{n}\big(M_X(0)\big) \;=\; \ell\mathrm{n}(1) \;=\; 0 \tag{3}$$

and

$$\lim_{y\to 0} M_X'(ty) \;=\; M_X'(0) \;=\; \mathrm{E}[X] \;=\; \mu \;=\; 0, \tag{4}$$

where the last equality is from our standardization assumption.

So after all of this build-up, we have

$$
\begin{aligned}
&\lim_{y\to 0} \frac{\ell\mathrm{n}\big(M_X(ty)\big)}{y^2} \\
&= \lim_{y\to 0} \frac{t\,M_X'(ty)}{2y\,M_X(ty)} \quad \text{(by (3) et L'Hôspital to deal with } 0/0\text{)} \\
&= \lim_{y\to 0} \frac{t^2\,M_X''(ty)}{2\,M_X(ty) + 2yt\,M_X'(ty)} \quad \text{(by (4) et L'Hôspital encore)} \\
&= \frac{t^2\,M_X''(0)}{2\,M_X(0)+0} = \frac{t^2\,\mathrm{E}[X^2]}{2} = \frac{t^2}{2} \quad (\mathrm{E}[X^2] = \sigma^2 - \mu^2 = 1).
\end{aligned}
$$

Remarks: We have a lot to say about such an important theorem.

(1) If n is large, then $\bar{X} \approx \text{Nor}(\mu, \sigma^2/n)$.

(2) The X_i's **don't have to be normal** for the CLT to work! It even works on discrete distributions!

(3) You usually need $n \geq 30$ observations for the approximation to work well. (Need fewer observations if the X_i's come from a symmetric distribution.)

(4) You can almost always use the CLT if the observations are iid.

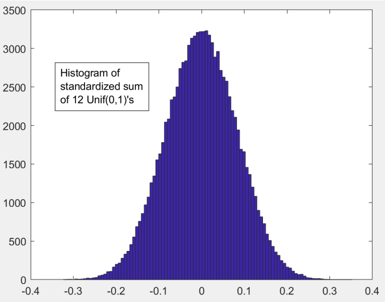
(5) In fact, there are versions of the CLT that are actually a lot more general than the theorem presented here!

Lesson 1.11 — Central Limit Theorem Examples

Example: To show that the Central Limit Theory really works, let's add up just $n = 12$ iid Unif(0,1)'s, $U_1, \ldots, U_n$. Let $S_n = \sum_{i=1}^{n} U_i$. Note that $\mathrm{E}[S_n] = n\mathrm{E}[U_i] = n/2$, and $\mathrm{Var}(S_n) = n\mathrm{Var}(U_i) = n/12$. Therefore,

$$Z_n \equiv \frac{S_n - n/2}{\sqrt{n/12}} = S_n - 6 \approx \mathrm{Nor}(0,1).$$

The histogram was compiled using 100,000 simulations of Z_{12}. It works! $\square$



Example: Suppose $X_1, \ldots, X_{100} \overset{\text{iid}}{\sim} \text{Exp}(1/1000)$.
Find $P(950 \le \bar{X} \le 1050)$.

Solution: Recall that if $X_i \sim \text{Exp}(\lambda)$, then $\text{E}[X_i] = 1/\lambda$ and $\text{Var}(X_i) = 1/\lambda^2$.

Further, if $\bar{X}$ is the sample mean based on n observations, then

$$\begin{aligned} \text{E}[\bar{X}] &= \text{E}[X_i] = 1/\lambda \quad \text{and} \\ \text{Var}(\bar{X}) &= \text{Var}(X_i)/n = 1/(n\lambda^2). \end{aligned}$$

For our problem, $\lambda = 1/1000$ and $n = 100$, so that $\text{E}[\bar{X}] = 1000$ and $\text{Var}(\bar{X}) = 10000$.

Georgia Tech

So by the CLT,

$$
\begin{aligned}
&P(950 \le \bar{X} \le 1050) \\
&\quad = \quad P\left(\frac{950 - \mathrm{E}[\bar{X}]}{\sqrt{\mathrm{Var}(\bar{X})}} \le \frac{\bar{X} - \mathrm{E}[\bar{X}]}{\sqrt{\mathrm{Var}(\bar{X})}} \le \frac{1050 - \mathrm{E}[\bar{X}]}{\sqrt{\mathrm{Var}(\bar{X})}}\right) \\
&\quad \doteq \quad P\left(\frac{950 - 1000}{100} \le Z \le \frac{1050 - 1000}{100}\right) \\
&\quad = \quad P\left(-\frac{1}{2} \le Z \le \frac{1}{2}\right) \;=\; 2\Phi(1/2) - 1 \;=\; 0.383. \quad \square
\end{aligned}
$$

Remark: This problem can be solved exactly if we have access to the Excel Erlang cdf function `GAMMADIST`. And what do you know, you end up with exactly the same answer of 0.383!

Example: Suppose $X_1, \ldots, X_{100}$ are iid from some distribution with mean 1000 and standard deviation 1000. Find $P(950 \leq \bar{X} \leq 1050)$.

Solution: By exactly the same manipulations as in the previous example, the answer $\doteq 0.383$.

Notice that we didn't care whether or not the data came from an exponential distribution. We just needed the mean and variance. $\Box$

Normal Approximation to the Binomial

Suppose $Y \sim \text{Bin}(n, p)$, where n is very large. In such cases, we usually approximate the Binomial via an appropriate Normal distribution.

The CLT applies since $Y = \sum_{i=1}^{n} X_i$, where the X_i's are iid Bern(p).

Then

$$\frac{Y - \text{E}[Y]}{\sqrt{\text{Var}(Y)}} = \frac{Y - np}{\sqrt{npq}} \approx \text{Nor}(0, 1).$$

The usual rule of thumb for the Normal approximation to the Binomial is that it works pretty well as long as $np \geq 5$ and $nq \geq 5$.

Why do we need such an approximation?

Example: Suppose $Y \sim \text{Bin}(100, 0.8)$ and we want

$$P(Y \geq 84) = \sum_{i=84}^{100} \binom{100}{i} (0.8)^i (0.2)^{100-i}.$$

Good luck with the binomial coefficients (they're too big) and the number of terms to sum up (it's going to get tedious). I'll come back to visit you in an hour.

The next example shows how to use the approximation.

Note that it incorporates a **"continuity correction"** to account for the fact that the Binomial is *discrete* while the Normal is *continuous*. If you don't want to use it, don't worry too much.

Example: The Braves play 100 independent baseball games, each of which they have probability 0.8 of winning. What's the probability that they win $\geq$ 84?

$Y \sim \text{Bin}(100, 0.8)$ and we want $P(Y \geq 84)$ (as in the last example)...

$$P(Y \geq 84) \;=\; P(Y \geq 83.5) \quad \text{("continuity correction")}$$

$$\doteq \; P\left(Z \geq \frac{83.5 - np}{\sqrt{npq}}\right) \quad \text{(CLT)}$$

$$= \; P\left(Z \geq \frac{83.5 - 80}{\sqrt{16}}\right) \;=\; P(Z \geq 0.875) \;=\; 0.1908.$$

The actual answer (using the true Bin(100,0.8) distribution) turns out to be 0.1923 — pretty close! □

Georgia Tech

Lesson 1.12 — Extensions — Multivariate Normal Distribution

Definition: (X, Y) has the **Bivariate Normal distribution** if it has pdf

$$f(x,y) \;=\; C \exp\left\{ \frac{-\left[z_X^2(x) - 2\rho z_X(x) z_Y(y) + z_Y^2(y) \right]}{2(1-\rho^2)} \right\},$$

where

$$\rho \;\equiv\; \mathrm{Corr}(X,Y), \quad C \;\equiv\; \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}},$$

$$z_X(x) \;\equiv\; \frac{x-\mu_X}{\sigma_X}, \quad \text{and} \quad z_Y(y) \;\equiv\; \frac{y-\mu_Y}{\sigma_Y}.$$

Pretty nasty joint pdf, eh?

In fact, $X \sim \text{Nor}(\mu_X, \sigma_X^2)$ and $Y \sim \text{Nor}(\mu_Y, \sigma_Y^2)$.

Example: (X, Y) could be a person's (height,weight). The two quantities are marginally normal, but positively correlated.

If you want to calculate bivariate normal probabilities, you'll need to evaluate quantities like

$$P(a < X < b,\, c < Y < d) \;=\; \int_c^d \int_a^b f(x, y)\, dx\, dy,$$

which will probably require numerical integration techniques.

Fun Fact (which will come up later when we discuss regression): The conditional distribution of Y given that $X = x$ is also normal. In particular,

$$Y|X = x \ \sim \ \text{Nor}\Big(\mu_Y + \rho(\sigma_Y/\sigma_X)(x - \mu_X),\ \sigma_Y^2(1-\rho^2)\Big).$$

Information about X helps to update the distribution of Y.

Example: Consider students at a university. Let X be their combined SAT scores (Math and Verbal), and Y their freshman GPA (out of 4).
Suppose a study reveals that

$$\mu_X = 1300, \quad \mu_Y = 2.3,$$

$$\sigma_X^2 = 6400, \quad \sigma_Y^2 = 0.25, \quad \rho = 0.6.$$

Find $P(Y \geq 2|X = 900)$.

First,

$$\begin{aligned} \mathrm{E}[Y|X=900] &= \mu_Y + \rho(\sigma_Y/\sigma_X)(x-\mu_X) \\ &= 2.3 + (0.6)(\sqrt{0.25/6400})(900-1300) \;=\; 0.8, \end{aligned}$$

indicating that the expected GPA of a kid with 900 SAT's will be 0.8.

Second,

$$\mathrm{Var}(Y|X=900) \;=\; \sigma_Y^2(1-\rho^2) \;=\; 0.16.$$

Thus,

$$Y|X=900 \;\sim\; \mathrm{Nor}(0.8, 0.16).$$

Now we can calculate

$$P(Y \geq 2|X=900) \;=\; P\left(Z \geq \frac{2-0.8}{\sqrt{0.16}}\right) \;=\; 1-\Phi(3) \;=\; 0.0013.$$

This guy doesn't have much chance of having a good GPA. $\square$

The bivariate normal distribution is easily generalized to the multivariate case.

Honors Definition: The random vector $\boldsymbol{X} = (X_1, \ldots, X_k)^{\mathrm{T}}$ has the **multivariate normal distribution** with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^{\mathrm{T}}$ and $k \times k$ **covariance matrix** $\boldsymbol{\Sigma} = (\sigma_{ij})$ if it has multivariate pdf

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right\}, \quad \boldsymbol{x} \in \mathbb{R}^k,$$

where $|\boldsymbol{\Sigma}|$ and $\boldsymbol{\Sigma}^{-1}$ are the determinant and inverse of $\boldsymbol{\Sigma}$, respectively.

It turns out that

$$\mathrm{E}[X_i] = \mu_i, \quad \mathrm{Var}(X_i) = \sigma_{ii}, \quad \mathrm{Cov}(X_i, X_j) = \sigma_{ij}.$$

Notation: $\boldsymbol{X} \sim \mathrm{Nor}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Lesson 1.13 — Extensions — Lognormal Distribution

Definition: If $Y \sim \text{Nor}(\mu_Y, \sigma_Y^2)$, then $X \equiv e^Y$ has the **Lognormal distribution** with parameters (μ_Y, σ_Y^2). This distribution has tremendous uses, e.g., in the pricing of certain stock options.

Turns Out: The pdf and moments of the lognormal are

$$
\begin{aligned}
f(x) &= \frac{1}{x\sigma_Y\sqrt{2\pi}}\,\exp\left\{-\frac{1}{2\sigma_Y^2}[\ell\text{n}(x) - \mu_Y]^2\right\},\ x > 0, \\
\text{E}[X] &= \exp\left\{\mu_Y + \frac{\sigma_Y^2}{2}\right\}, \\
\text{Var}(X) &= \exp\left\{2\mu_Y + \sigma_Y^2\right\}\big(\exp\left\{\sigma_Y^2\right\} - 1\big).
\end{aligned}
$$

Example: Suppose $Y \sim \text{Nor}(10, 4)$ and let $X = e^Y$. Then

$$\begin{aligned} P(X \le 1000) &= P\big(Y \le \ell\text{n}(1000)\big) \\ &= P\left(Z \le \frac{\ell\text{n}(1000) - 10}{2}\right) \\ &= \Phi(-1.55) \;=\; 0.061. \quad \Box \end{aligned}$$

Honors Example (How to Win a Nobel Prize:) It is well-known that stock prices are closely related to the lognormal distribution. In fact, it's common to use the following model for a stock price at a fixed time t,

$$S(t) \;=\; S(0) \exp\left\{ \left(\mu - \frac{\sigma^2}{2}\right) t + \sigma\sqrt{t} Z \right\}, \quad t \ge 0,$$

where μ is related to the "drift" of the stock price (i.e., the natural rate of increase), σ is its "volatility" (how much the stock bounces around), $S(0)$ is the initial price, and Z is a standard normal RV.

An active area of finance is to estimate **option prices**. For example, a so-called **European call option** C permits its owner, who pays an up-front fee for the privilege, to purchase the stock at a pre-agreed **strike price** k, at a pre-determined **expiry date** T.

For instance, suppose IBM is currently selling for \$100 a share. If I think that the stock will go up in value, I may want to pay \$3/share now for the right to buy IBM at \$105 three months from now.

- If IBM is worth \$120 three months from now, I'll be able to buy it for only \$105, and will have made a profit of $\$120 - \$105 - \$3 = \$12$.
- If IBM is selling for \$107 three months hence, I can still buy it for \$105, and will lose \$1 (recouping \$2 from my original option purchase).
- If IBM is selling for \$95, then I won't exercise my option, and will walk away with my tail between my legs having lost my original \$3.

So what is the option worth (and what should I pay for it)? Its expected value is given by

$$\mathrm{E}[C] \; = \; e^{-rT}\mathrm{E}\big[\big(S(T)-k\big)^{+}\big],$$

where

- $x^{+} \equiv \max\{0, x\}$.
- r, the "risk-free" interest rate (e.g., what you can get from a U.S. Treasury bond). This is used instead of the drift μ.
- The term e^{-rT} denotes the time-value of money, i.e., a depreciation term corresponding to the interest I could've made had I used my money to buy a Treasury note.

Black and Scholes won a Nobel Prize for calculating $\mathrm{E}[C]$. We'll get the same answer via a different method — but, alas, no Nobel Prize. ☹

$$
\begin{aligned}
\mathrm{E}[C] &= e^{-rT}\mathrm{E}\left[S(0)\exp\left\{\left(r-\frac{\sigma^2}{2}\right)T+\sigma\sqrt{T}Z\right\}-k\right]^+ \\
&= e^{-rT}\int_{-\infty}^{\infty}\left[S(0)\exp\left\{\left(r-\frac{\sigma^2}{2}\right)T+\sigma\sqrt{T}z\right\}-k\right]^+\phi(z)\,dz \\
&\qquad \text{(via the standard conditioning argument)} \\
&= S(0)\Phi(b+\sigma\sqrt{T})-ke^{-rT}\Phi(b) \quad \text{(after lots of algebra)},
\end{aligned}
$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the Nor(0,1) pdf and cdf, respectively, and

$$
b \equiv \frac{rT-\frac{\sigma^2T}{2}-\ell\mathrm{n}(k/S(0))}{\sigma\sqrt{T}}.
$$

There are many generalizations of this problem that are used in practical finance problems, but this is the starting point. Meanwhile, get your tickets to Norway or Sweden or wherever they give out the Nobel! ☺

Lesson 1.14 — Computer Stuff

Evaluating pmf's / pdf's and cdf's

We can use various computer packages such as Excel, Minitab, R, SAS, etc., to calculate pmf's / pdf's and cdf's for a variety of common distributions. For instance, in Excel, we find the functions:

`BINOMDIST` = Binomial distribution
`EXPONDIST` = Exponential
`NEGBINOMDIST` = Negative Binomial
`NORMDIST` and `NORMSDIST` = Normal and Standard Normal
`POISSON` = Poisson

Functions such as `NORMSINV` and `TINV` can calculate the inverses of the standard normal and t distributions, respectively.

Simulating Random Variables

Motivation: Simulations are used to evaluate a variety of real-world processes that contain inherent randomness, e.g., queueing systems, inventory systems, manufacturing systems, etc. In order to run simulations, you need to generate various RVs, e.g., arrival times, service times, failure times, etc.

Examples: There are numerous ways to simulate RVs.

- The Excel function `RAND` simulates a Unif(0,1) RV.
- If $U_1, U_2 \overset{\text{iid}}{\sim}$ Unif(0,1), then $U_1 + U_2$ is Triangular(0,1,2). This is simply `RAND()+RAND()` in Excel.
- The Inverse Transform Theorem gives $(-1/\lambda)\ell\text{n}(U) \sim \text{Exp}(\lambda)$.
- If $U_1, U_2, \ldots, U_k \overset{\text{iid}}{\sim}$ Unif(0,1), then $\sum_{i=1}^k (-1/\lambda)\ell\text{n}(U_i)$ is Erlang$_k(\lambda)$ because it is the sum of iid Exp(λ)'s.
- It can be shown that $X = \lceil \ell\text{n}(U)/\ell\text{n}(1-p) \rceil \sim \text{Geom}(p)$, where $\lceil \cdot \rceil$ is the "ceiling" (integer round-up) function.

The simulation of RVs is actually the topic of another course, but we will end this module with a remarkable method for generating normal RVs.

Theorem (Box and Muller): If $U_1, U_2 \stackrel{\text{iid}}{\sim} \text{Unif}(0,1)$, then

$$Z_1 = \sqrt{-2\ell\text{n}(U_1)}\cos(2\pi U_2) \quad \text{and} \quad Z_2 = \sqrt{-2\ell\text{n}(U_1)}\sin(2\pi U_2)$$

are iid Nor(0,1).

Example: Suppose that $U_1 = 0.3$ and $U_2 = 0.8$ are realizations of two iid Unif(0,1)'s. Box–Muller gives the following two iid standard normals.

$$\begin{aligned} Z_1 &= \sqrt{-2\ell\text{n}(U_1)}\cos(2\pi U_2) = 0.480 \\ Z_2 &= \sqrt{-2\ell\text{n}(U_1)}\sin(2\pi U_2) = -1.476. \quad \Box \end{aligned}$$

Georgia Tech

Remarks:

- There are many other ways to generate Nor(0,1)'s, but this is the perhaps the easiest.
- It is essential that the cosine and sine must be calculated in *radians*, not degrees.
- To get $X \sim \text{Nor}(\mu, \sigma^2)$ from $Z \sim \text{Nor}(0, 1)$, just take $X = \mu + \sigma Z$.
- Amazingly, it's "Muller", not "Müller".
  See https://www.youtube.com/watch?v=nntGTK2Fhb0

Honors Proof: We follow the method from Module 3 to calculate the joint pdf of a function of two RVs. Namely, if we can express $U_1 = k_1(Z_1, Z_2)$ and $U_2 = k_2(Z_1, Z_2)$ for some functions $k_1(\cdot,\cdot)$ and $k_2(\cdot,\cdot)$, then the joint pdf of (Z_1, Z_2) is given by

$$\begin{aligned} g(z_1, z_2) &= f_{U_1}\big(k_1(z_1, z_2)\big) f_{U_2}\big(k_2(z_1, z_2)\big) \left| \frac{\partial u_1}{\partial z_1}\frac{\partial u_2}{\partial z_2} - \frac{\partial u_2}{\partial z_1}\frac{\partial u_1}{\partial z_2} \right| \\ &= \left| \frac{\partial u_1}{\partial z_1}\frac{\partial u_2}{\partial z_2} - \frac{\partial u_2}{\partial z_1}\frac{\partial u_1}{\partial z_2} \right| \quad \big(U_1 \text{ and } U_2 \text{ are iid Unif(0,1)}\big). \end{aligned}$$

In order to obtain the functions $k_1(Z_1, Z_2)$ and $k_2(Z_1, Z_2)$, note that

$$Z_1^2 + Z_2^2 \;=\; -2\,\ell\text{n}(U_1)\big[\cos^2(2\pi U_2) + \sin^2(2\pi U_2)\big] \;=\; -2\,\ell\text{n}(U_1),$$

so that

$$U_1 \;=\; e^{-(Z_1^2 + Z_2^2)/2}.$$

This immediately implies that

$$\begin{aligned} Z_1^2 &= -2\,\ell\mathrm{n}(U_1)\,\cos^2(2\pi U_2) \\ &= -2\,\ell\mathrm{n}\big(e^{-(Z_1^2+Z_2^2)/2}\big)\,\cos^2(2\pi U_2) \\ &= (Z_1^2+Z_2^2)\,\cos^2(2\pi U_2), \end{aligned}$$

so that

$$U_2 = \frac{1}{2\pi}\arccos\left(\pm\sqrt{\frac{Z_1^2}{Z_1^2+Z_2^2}}\right) = \frac{1}{2\pi}\arccos\left(\sqrt{\frac{Z_1^2}{Z_1^2+Z_2^2}}\right),$$

where we (non-rigorously) get rid of the "$\pm$" to balance off the fact that the range of $y = \arccos(x)$ is only regarded to be $0 \le y \le \pi$ (not $0 \le y \le 2\pi$).

Now some derivative fun:

$$\frac{\partial u_1}{\partial z_i} = \frac{\partial}{\partial z_i} e^{-(z_1^2+z_2^2)/2} = -z_i\, e^{-(z_1^2+z_2^2)/2}, \; i = 1, 2,$$

$$\begin{aligned}
\frac{\partial u_2}{\partial z_1} &= \frac{\partial}{\partial z_1} \frac{1}{2\pi} \arccos\left(\sqrt{\frac{z_1^2}{z_1^2+z_2^2}}\right) \\
&= \frac{1}{2\pi} \frac{-1}{\sqrt{1-\frac{z_1^2}{z_1^2+z_2^2}}} \frac{\partial}{\partial z_1} \sqrt{\frac{z_1^2}{z_1^2+z_2^2}} \quad \text{(chain rule)} \\
&= \frac{1}{2\pi} \frac{-1}{\sqrt{\frac{z_2^2}{z_1^2+z_2^2}}} \frac{1}{2} \left(\frac{z_1^2}{z_1^2+z_2^2}\right)^{-1/2} \frac{\partial}{\partial z_1} \frac{z_1^2}{z_1^2+z_2^2} \quad \text{(chain rule again)} \\
&= \frac{-(z_1^2+z_2^2)}{4\pi z_1 z_2} \frac{2 z_1 z_2^2}{(z_1^2+z_2^2)^2} = \frac{-z_2}{2\pi(z_1^2+z_2^2)}, \quad \text{and} \\
\frac{\partial u_2}{\partial z_2} &= \frac{z_1}{2\pi(z_1^2+z_2^2)} \quad \text{(after similar algebra).}
\end{aligned}$$

Then we finally have

$$\begin{aligned}
g(z_1, z_2) &= \left| \frac{\partial u_1}{\partial z_1} \frac{\partial u_2}{\partial z_2} - \frac{\partial u_2}{\partial z_1} \frac{\partial u_1}{\partial z_2} \right| \\
&= \left| -z_1 \, e^{-(z_1^2+z_2^2)/2} \frac{z_1}{2\pi(z_1^2+z_2^2)} - \frac{z_2}{2\pi(z_1^2+z_2^2)} \, z_2 \, e^{-(z_1^2+z_2^2)/2} \right| \\
&= \frac{1}{2\pi} \, e^{-(z_1^2+z_2^2)/2} = \left(\frac{1}{\sqrt{2\pi}} \, e^{-z_1^2/2} \right) \left(\frac{1}{\sqrt{2\pi}} \, e^{-z_2^2/2} \right),
\end{aligned}$$

which is the product of two iid Nor(0,1) pdf's, so we are done! ☺