3. Distributions

Dave Goldsman

Georgia Institute of Technology, Atlanta, GA, USA

6/10/14
1. Discrete Distributions
   - Bernoulli and Binomial Distributions
   - Hypergeometric Distribution
   - Geometric and Negative Binomial Distributions
   - Poisson Distribution

2. Continuous Distributions
   - Uniform Distribution
   - Exponential, Erlang, and Gamma Distributions
   - Other Continuous Distributions

3. Normal Distribution
   - Basics
   - Standard Normal Distribution
   - Sample Mean of Normal Observations
   - Central Limit Theorem
   - Extensions of the Normal Distribution

4. Computer Stuff
Bernoulli($p$) and Binomial($n, p$) Distributions

The Bern($p$) distribution is given by

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

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

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

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

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

$$P(Y = 4) = \binom{5}{4} \left(\frac{1}{6}\right)^4 \left(\frac{5}{6}\right)^{5-4}.$$
\( Y \sim \text{Bin}(n, p) \) implies

\[
E[Y] = E\left[ \sum_{i=1}^{n} X_i \right] = \sum_{i=1}^{n} E[X_i] = np
\]

and, similarly,

\[
\text{Var}(Y) = npq.
\]

We’ve already seen that \( M_Y(t) = (pe^t + q)^n \).

Binomials add up: If \( Y_1, \ldots, Y_k \) are indep and \( Y_i \sim \text{Bin}(n_i, p) \), then

\[
\sum_{i=1}^{k} Y_i \sim \text{Bin}\left( \sum_{i=1}^{k} n_i, p \right).
\]
Hypergeometric Distribution

You have $a$ objects of type 1 and $b$ objects of type 2.

Select $n$ objects w/o replacement from the $a + b$.

Let $X$ be the number of type 1’s selected.

$$P(X = k) = \frac{\binom{a}{k} \binom{b}{n-k}}{\binom{a+b}{n}}, \quad k = 0, 1, \ldots, \min(a, n).$$
After some algebra, it turns out that

\[ E[X] = n \left( \frac{a}{a + b} \right) \quad \text{and} \quad \text{Var}(X) = n \left( \frac{a}{a + b} \right) \left( 1 - \frac{a}{a + b} \right) \left( \frac{a + b - n}{a + b - 1} \right). \]

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

\[ P(\text{exactly 3 reds are picked}) = \frac{\binom{15}{3} \binom{10}{4}}{\binom{25}{7}} \]
Definition: Suppose we consider an infinite sequence of indep Bern\((p)\) trials.

Let \(Z\) equal the number of trials \textit{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\((p)\) distribution.
The mgf of the Geom(p) is

\[ M_Z(t) = \mathbb{E}[e^{tZ}] = \sum_{k=1}^{\infty} e^{tk} q^{k-1} p \]

\[ = p e^t \sum_{k=0}^{\infty} (q e^t)^k \]

\[ = \frac{p e^t}{1 - q e^t}, \quad \text{for } q e^t < 1. \]

So

\[ M_Z(t) = \frac{p e^t}{1 - q e^t}, \quad \text{for } t < \ln(1/q). \]
Thus,

\[ E[Z] = \left. \frac{d}{dt} M_Z(t)\right|_{t=0} = \left. \frac{(1 - qe^t)(pe^t) - (-qe^t)(pe^t)}{(1 - qe^t)^2} \right|_{t=0} = \frac{pe^t}{(1 - qe^t)^2} \bigg|_{t=0} = \frac{p}{(1-q)^2} = \frac{1}{p}. \]

(We can also prove this directly from the definition of expected value.)

Similarly, after a lot of algebra,

\[ E[Z^2] = \left. \frac{d^2}{dt^2} M_Z(t)\right|_{t=0} = \frac{2 - p}{p^2}, \]

so that

\[ \text{Var}(Z) = E[Z^2] - (E[Z])^2 = \frac{2 - p}{p^2} - \frac{1}{p^2} = \frac{q}{p^2}. \]
Example: Toss a die repeatedly. What’s the prob that we observe a ‘3’ for the first time on the 8th toss?

Answer: The number of tosses we need is \( Z \sim \text{Geom}(1/6) \).
\[
P(Z = 8) = (5/6)^7 (1/6).
\]

How many tosses would we expect to take?

Answer: \( E[Z] = 1/p = 6 \) tosses.
Memoryless Property of Geometric

Theorem: Suppose that $Z \sim \text{Geom}(p)$. 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 prob that it will occur after an additional $t$ time units is the same as the (unconditional) prob that it will occur after time $t$ — it forgot that it made it past time $s$!
Proof: First of all, for any \( t = 0, 1, 2, \ldots \),

\[
P(Z > t) = \sum_{j=t+1}^{\infty} q^{j-1}p = pq^t \sum_{j=0}^{\infty} q^j = \frac{pq^t}{1-q} = q^t. \quad (\ast)
\]

In addition,

\[
P(Z > s + t|Z > s) = \frac{P(Z > s + t \cap Z > s)}{P(Z > s)}
\]

\[
= \frac{P(Z > s + t)}{P(Z > s)}
\]

\[
= \frac{q^{s+t}}{q^s} \quad \text{by (\ast)}
\]

\[
= q^t.
\]

Thus, \( P(Z > s + t|Z > s) = P(Z > t) \). Done.
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 prob that we’ll need more than 2 more tosses before we observe a 5?

Let \( Z \) be the number of tosses required. By the memoryless prop (with \( s = 4 \) and \( t = 2 \)) and (*), we want

\[
P(Z > 6 | Z > 4) = P(Z > 2) = (5/6)^2.
\]

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 indep Bern($p$) trials.

Now let $W$ equal the number of trials until the $r$th 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 $r$th success at time $k$.

We say that $W$ has the **Negative Binomial($r, p$) distribution** (aka the Pascal distrn).

Example: ‘FFFFFSFS’ 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 $Z_1, \ldots, Z_r \overset{iid}{\sim} \text{Geom}(p)$, then $W = \sum_{i=1}^{n} Z_i \sim \text{NegBin}(r, p)$.

Proof: Won’t do it here, but you can use the mgf technique.

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 $i$th success.

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

$$E[W] = rE[Z_i] = r/p,$$

$$\text{Var}(W) = r\text{Var}(Z_i) = rq/p^2,$$

$$M_W(t) = [M_{Z_i}(t)]^r = \left(\frac{pe^t}{1 - qe^t}\right)^r.$$
Just to be complete, let’s get the pmf of $W$. Note that $W = k$ iff get exactly $r - 1$ successes by time $k - 1$, and then the $r$th 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 prob that we’ll need exactly 7 tosses?

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

$$P(W = 7) = \binom{7-1}{3-1} (1/6)^3 (5/6)^7 - 3.$$
How are the Binomial and NegBin Related?

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

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

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

\[ \mathbb{E}[W] = r/p, \quad \text{Var}(W) = rq/p^2. \]
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 (time).
(b) Defects on a wire (length).
(c) Raisins in cookie dough (volume).

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

In the above examples, we might have:
(a) $\lambda = 10/\text{min.}$  (b) $\lambda = 0.5/\text{ft.}$  (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** is one that satisfies the following:

1. There is a short enough interval of time \( h \), such that, for all \( t \),
   \[
   P(N(t + h) - N(t) = 0) = 1 - \lambda h + o(h)
   \]
   \[
   P(N(t + h) - N(t) = 1) = \lambda h + o(h)
   \]
   \[
   P(N(t + h) - N(t) \geq 2) = o(h)
   \]

2. The distribution of the “increment” \( N(t + h) - N(t) \) only depends on the length \( h \).

3. If \( a < b < c < d \), then the two “increments” \( N(d) - N(c) \) and \( N(b) - N(a) \) are *indep* RV’s.
English translation of Poisson process assumptions.

(1) Arrivals basically occur one-at-a-time, and then at rate $\lambda$/unit time. (We must make sure that $\lambda$ doesn’t change over time.)

(2) The arrival pattern is stationary — it doesn’t change over time.

(3) The numbers of arrivals in two disjoint time intervals are indep.

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 indep of each other.

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 indep. This ain’t a Poisson process.
Definition: Let $X$ be the number of occurrences in a Poisson($\lambda$) process in a *unit interval* of time. Then $X$ has the **Poisson distribution** with parameter $\lambda$.

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

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

Remark: The value of $\lambda$ can be changed simply by changing the units of time.

Example:
$X = \# \text{ calls to a switchboard in 1 minute} \sim \text{Pois}(3)$
$Y = \# \text{ calls to a switchboard in 5 minutes} \sim \text{Pois}(15)$
$Z = \# \text{ calls to a switchboard in 10 sec} \sim \text{Pois}(0.5)$
Theorem: $X \sim \text{Pois}(\lambda) \Rightarrow \text{mgf is } M_X(t) = e^{\lambda(e^t - 1)}.$

Proof:

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

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

Proof (using mgf):

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

Similarly, 
\[
E[X^2] = \left. \frac{d^2}{dt^2} M_X(t) \right|_{t=0} = \left. \frac{d}{dt} \left( \frac{d}{dt} M_X(t) \right) \right|_{t=0}
\]
\[
= \lambda \left. \frac{d}{dt} (e^t M_X(t)) \right|_{t=0}
\]
\[
= \lambda \left[ e^t M_X(t) + e^t \frac{d}{dt} M_X(t) \right] \bigg|_{t=0}
\]
\[
= \lambda e^t \left[ M_X(t) + \lambda e^t M_X(t) \right] \bigg|_{t=0} = \lambda (1 + \lambda).
\]

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

Example: Calls to a switchboard arrive as a Poisson process with rate 3 calls/min. Let \( X = \) number of calls in 40 sec.

So \( X \sim \text{Pois}(2), \ E[X] = \text{Var}(X) = 2, \ P(X \leq 3) = \sum_{k=0}^{3} e^{-2} \frac{2^k}{k!}. \)
Theorem (Additive Property of Poissons): Suppose $X_1, \ldots, X_n$ are *indep* with $X_i \sim \text{Pois}(\lambda_i)$, $i = 1, \ldots, n$. Then

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

Proof: Since the $X_i$’s are indep, 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 $\text{Pois}(\sum_{i=1}^{n} \lambda_i)$ distribution.

Example: Cars driven by males [females] arrive at a parking lot according to a Poisson process with a rate of 3/hr [5/hr]. All arrivals are indep. What’s the probability of exactly 2 arrivals in the next 30 min?

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


Outline

1. Discrete Distributions
   - Bernoulli and Binomial Distributions
   - Hypergeometric Distribution
   - Geometric and Negative Binomial Distributions
   - Poisson Distribution

2. Continuous Distributions
   - Uniform Distribution
   - Exponential, Erlang, and Gamma Distributions
   - Other Continuous Distributions

3. Normal Distribution
   - Basics
   - Standard Normal Distribution
   - Sample Mean of Normal Observations
   - Central Limit Theorem
   - Extensions of the Normal Distribution

4. Computer Stuff
Uniform \((a, b)\) Distribution

Recall that

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

Previous work showed that

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

We can also derive the mgf,

\[
M_X(t) = E[e^{tX}] = \int_a^b e^{tx} \frac{1}{b-a} \, dx = \frac{e^{tb} - e^{ta}}{t(b-a)}.
\]
Recall that the **Exponential**$(\lambda)$ **distribution** has pdf

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

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

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

We also derived the mgf,

\[
M_X(t) = E[e^{tX}] = \int_0^\infty e^{tx} f(x) \, dx = \frac{\lambda}{\lambda - t}, \quad t < \lambda.
\]
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 prob that \( X \) will survive an additional \( t \) time units is the (unconditional) prob that it will survive at least \( t \) — it forgot that it made it past time \( s \)!

Proof:

\[
P(X > s + t | X > s) = \frac{P(X > s + t \cap X > s)}{P(X > s)} = \frac{P(X > s + t)}{P(X > s)}
\]

\[
= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X > t).
\]
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 prob that it’ll survive another 10?

\[ P(X > 30|X > 20) = P(X > 10) = e^{-\lambda x} = e^{-(1/10)(10)} = e^{-1}. \]

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.

Remark: The exponential is the *only* cts distrn with the memoryless property.

Remark: Look at \( E[X] \) and \( \text{Var}(X) \) for the Geometric distrn 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 $\lambda$ — always good as new! This is clearly a result of the memoryless property.
The Exponential is also related to the Poisson!

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

Proof:

\[
F(x) = P(X \leq x) = 1 - P(\text{no arrls in } [0, x]) = 1 - e^{-\lambda x}^\text{0} (\text{since } \# \text{ arrls in } [0, x] \text{ is Pois}(\lambda x))
\]

\[
= 1 - e^{-\lambda x}.
\]

Amazingly, it can be shown (after a lot of work) that the interarrival times of a Poisson process are all i.i.d. \( \text{Exp}(\lambda) \)! See for yourself when you take a stochastic processes course.
Example: Suppose that arrivals to a shopping center are from a Poisson process 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 custs 13 and 14 be $X$. Since we have a Poisson process, the interarrivals are iid $\text{Exp}(\lambda = 20$/hr), so

$$P(X > 4 \text{ min}) = P(X > 1/15 \text{ hr}) = e^{-\lambda t} = e^{-20/15}.$$
Definition: Suppose $X_1, \ldots, X_k \overset{iid}{\sim} \text{Exp}(\lambda)$, and let $S = \sum_{i=1}^{k} X_i$. Then $S$ has the **Erlang** _k_ distribution with parameter $\lambda$.

The Erlang is simply the sum of i.i.d. exponentials.

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

The pdf and cdf of the Erlang are

$$f(s) = \frac{\lambda^k e^{-\lambda s} s^{k-1}}{(k-1)!}, \quad s \geq 0,$$

$$F(s) = 1 - \sum_{i=0}^{k-1} \frac{e^{-\lambda s} (\lambda s)^i}{i!}.$$

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:

\[
E[S] = E\left[\sum_{i=1}^{k} X_i\right] = \sum_{i=1}^{k} E[X_i] = k/\lambda
\]

\[
\text{Var}(S) = k/\lambda^2
\]

\[
M_{S}(t) = \left(\frac{\lambda}{\lambda - t}\right)^k.
\]

Example: Suppose \(X\) and \(Y\) are i.i.d. \text{Exp}(2). Find \(P(X + Y < 1)\).

\[
P(X + Y < 1) = 1 - \sum_{i=0}^{k-1} \frac{e^{-\lambda s} (\lambda s)^i}{i!} = 1 - \sum_{i=0}^{2-1} \frac{e^{-(2 \cdot 1)} (2 \cdot 1)^i}{i!} = 0.594
\]
Definition: $X$ has the **gamma distribution** with parameters $\alpha > 0$ and $\lambda > 0$ if it has pdf

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

where

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

is the **gamma function**.

Remark: The gamma distrn generalizes the Erlang distrn (where $\alpha$ has to be a positive integer). It has the same expected value and variance as the Erlang, with $\alpha$ in place of $k$.

Remark: If $\alpha$ is a positive integer, then $\Gamma(\alpha) = (\alpha - 1)!$.

Party trick: $\Gamma(1/2) = \sqrt{\pi}$. 
**Triangular** \((a, b, c)\) **Distribution** — good for modeling RV’s on the basis of limited data (min, mode, max).

\[
f(x) = \begin{cases} 
\frac{2(x-a)}{(b-a)(c-a)} & a < x \leq b \\
\frac{2(c-x)}{(c-b)(c-a)} & b < x < c \\
0 & \text{otherwise}
\end{cases}
\]

\[
E[X] = \frac{a + b + c}{3} \quad \text{and} \quad \text{Var}(X) = \text{mess}
\]
Beta\((a, b)\) Distribution — good for modeling RV’s that are restricted to an interval.

\[
f(x) = \begin{cases} 
\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} & 0 < x < 1 \\
0 & \text{otherwise}
\end{cases}
\]

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

This distribution gets its name from the \textit{beta function}, which is defined as

\[
\beta(a, b) \equiv \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 x^{a-1}(1-x)^{b-1} \, dx.
\]
Weibull\((a, b)\) Distribution — good for modeling reliability models. \(a\) is the “scale” parameter, and \(b\) is the “shape” parameter.

\[
f(x) = \begin{cases} 
ab (ax)^{b-1} e^{-(ax)^b} & x > 0 \\
0 & \text{otherwise}
\end{cases}
\]

\[
F(x) = 1 - \exp[-(ax)^b], \quad x > 0.
\]

\[
E[X] = \frac{1}{a} \Gamma\left(1 + \frac{1}{b}\right) \quad \text{and} \quad \text{Var}(X) = \text{slight mess}
\]
Remark: The exponential is a special case of the Weibull.

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

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

The prob that it fails before 2000 hrs is

$$F(2000) = 1 - \exp\left[-\left(\frac{2000}{200}\right)^{1/3}\right] = 0.884.$$
Cauchy 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: \( X_1, \ldots, X_n \overset{iid}{\sim} \text{Cauchy} \Rightarrow \sum_{i=1}^{n} X_i/n \sim \text{Cauchy}. \) Even you take the average of a bunch of Cauchys, you’re right back where you started!
Alphabet Soup of Other Distrns

$\chi^2$ distribution — coming up in the statistics portion

t distribution — coming up

$F$ distribution — coming up

Pareto, LaPlace, Rayleigh, Gumbel distributions

Etc. . .
Normal Distribution

So important that we’ll give it an entire section.

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

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

\( f(x) \) is “bell-shaped” and symmetric around \( x = \mu \), with tails falling off quickly as you move away from \( \mu \).

Small \( \sigma^2 \) corresponds to a “tall, skinny” bell curve; large \( \sigma^2 \) gives a “short, fat” bell curve.

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

Examples: Heights, weights, SAT scores, crop yields, and averages of things tend to be normal.
Fun Fact (1): $\int_{\mathbb{R}} f(x) \, dx = 1$.

Proof: Transform to polar coordinates. Good luck.

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): $E[X] = \mu$ and $\text{Var}(X) = \sigma^2$.

Proof: Integration by parts or mgf (below).

Fun Fact (5): 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 indep with $X_i \sim \text{Nor}(\mu_i, \sigma_i^2), i = 1, \ldots, n$, then

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

So a linear combination of indep normals is itself normal.
Proof: Since \( Y \) is a linear function,

\[
M_Y(t) = M_{\sum_i a_i X_i + b}(t) = e^{tb} M_{\sum_i a_i X_i}(t)
\]

\[
= e^{tb} \prod_{i=1}^{n} M_{a_i X_i}(t) \quad (X_i \text{'s indep})
\]

\[
= e^{tb} \prod_{i=1}^{n} M_{X_i}(a_i t) \quad (mgf of linear fn)
\]

\[
= e^{tb} \prod_{i=1}^{n} \exp \left[ \mu_i (a_i t) + \frac{1}{2} \sigma_i^2 (a_i t)^2 \right] \quad (normal mgf)
\]

\[
= \exp \left[ \left( \sum_{i=1}^{n} \mu_i a_i + b \right) t + \frac{1}{2} \left( \sum_{i=1}^{n} a_i^2 \sigma_i^2 \right) t^2 \right] \quad (Done.)
\]
Remark: A normal distrn is *completely characterized* by its mean and variance.

By the above, we know that a linear combination of indep normals is still normal. Therefore, when we add up indep 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 $X, Y$ indep. Find the distrn of $2X - 3Y$.

Solution: This is *normal* with

\[
E[2X - 3Y] = 2E[X] - 3E[Y] = 2(3) - 3(4) = -6
\]

and

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

Thus, $2X - 3Y \sim \text{Nor}(-6, 70)$. 
Corollary (of Theorem):

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

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

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 Cor with \( a = 1/\sigma \) and \( b = -\mu/\sigma \).
Standard Normal Distribution

Definition: The $\text{Nor}(0, 1)$ distrn 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:

\[ P(Z \leq a) = \Phi(a) \]
\[ P(Z \geq b) = 1 - \Phi(b) \]
\[ P(a \leq Z \leq b) = \Phi(b) - \Phi(a) \]

\[ \Phi(0) = 1/2 \]
\[ \Phi(-b) = P(Z \leq -b) = P(Z \geq b) = 1 - \Phi(b) \]
\[ P(-b \leq Z \leq b) = \Phi(b) - \Phi(-b) = 2\Phi(b) - 1 \]

Then

\[ P(\mu - k\sigma \leq X \leq \mu + k\sigma) = P(-k \leq Z \leq k) = 2\Phi(k) - 1. \]

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. Or you can use software calls, like `NORMDIST` in Excel (which calculates the cdf for *any* normal distribution.)

<table>
<thead>
<tr>
<th>$z$</th>
<th>$\Phi(z) = P(Z \leq z)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.5000</td>
</tr>
<tr>
<td>1.00</td>
<td>0.8413</td>
</tr>
<tr>
<td>1.28</td>
<td>0.8997 $\approx 0.90$</td>
</tr>
<tr>
<td>1.645</td>
<td>0.9500</td>
</tr>
<tr>
<td>1.96</td>
<td>0.9750</td>
</tr>
<tr>
<td>2.33</td>
<td>0.9901 $\approx 0.99$</td>
</tr>
<tr>
<td>3.00</td>
<td>0.9987</td>
</tr>
<tr>
<td>4.00</td>
<td>$\approx 1.0000$</td>
</tr>
</tbody>
</table>
By the discussion on the last two pages, the probability that any normal RV is within $k$ standard deviations of its mean is

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

For $k = 1$, this probability is $2(0.8413) - 1 = 0.6826$.

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!
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 $\Phi(z) = p$.

<table>
<thead>
<tr>
<th>$p$</th>
<th>$\Phi^{-1}(p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.90</td>
<td>1.28</td>
</tr>
<tr>
<td>0.95</td>
<td>1.645</td>
</tr>
<tr>
<td>0.975</td>
<td>1.96</td>
</tr>
<tr>
<td>0.99</td>
<td>2.33</td>
</tr>
<tr>
<td>0.995</td>
<td>2.58</td>
</tr>
</tbody>
</table>
Example: $X \sim \text{Nor}(21, 4)$. Find $P(19 < X < 22.5)$. Standardizing, we get

$$P(19 < X < 22.5)$$

$$= P\left(\frac{19 - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{22.5 - \mu}{\sigma}\right)$$

$$= P\left(\frac{19 - 21}{2} < Z < \frac{22.5 - 21}{2}\right)$$

$$= P(-1 < Z < 0.75)$$

$$= \Phi(0.75) - \Phi(-1)$$

$$= \Phi(0.75) - [1 - \Phi(1)]$$

$$= 0.7734 - [1 - 0.8413] = 0.6147.$$
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 \emph{independently} at random.

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

Note that

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

Then

$$P(W > M) = P(W - M > 0)$$
$$= P\left(Z > \frac{0 + 3}{\sqrt{5}}\right)$$
$$= 1 - \Phi\left(3/\sqrt{5}\right)$$
$$\approx 1 - 0.91 = 0.09.$$
Sample Mean of Normal Observations

The *sample mean* of \( X_1, \ldots, X_n \) is

\[
\bar{X} \equiv \frac{\sum_{i=1}^{n} X_i}{n}.
\]

Corollary (of old Theorem):

\( X_1, \ldots, X_n \overset{iid}{\sim} \text{Nor}(\mu, \sigma^2) \Rightarrow \bar{X} \sim \text{Nor}(\mu, \sigma^2/n) \).

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

Remark: This result is *very significant!* As the number of observations increases, \( \text{Var}(\bar{X}) \) gets *smaller* (while \( \mathbb{E}[\bar{X}] \) remains constant).

In the upcoming statistics portion of the course, we’ll learn that this makes the sample mean \( \bar{X} \) an excellent *estimator* for the mean \( \mu \), which is typically unknown in practice.
Example: Suppose that \( X_1, \ldots, X_n \sim \text{Nor}(\mu, 16) \). Find the sample size \( n \) such that

\[
P(|\bar{X} - \mu| \leq 1) \geq 0.95.
\]

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

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

\[
P(|\bar{X} - \mu| \leq 1) = P(-1 \leq \bar{X} - \mu \leq 1)
\]

\[
= P\left(\frac{-1}{4/\sqrt{n}} \leq \frac{\bar{X} - \mu}{4/\sqrt{n}} \leq \frac{1}{4/\sqrt{n}}\right)
\]

\[
= P\left(\frac{-\sqrt{n}}{4} \leq Z \leq \frac{\sqrt{n}}{4}\right)
\]

\[
= 2\Phi(\sqrt{n}/4) - 1.
\]
Now we have to find $n$ such that this probability is at least 0.95. . . .

$$2\Phi(\sqrt{n}/4) - 1 \geq 0.95 \text{ iff } \Phi(\sqrt{n}/4) \geq 0.975 \text{ iff }$$

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

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 $\mu$. 
The most important theorem in prob and stats.

**Central Limit Theorem:** Suppose $X_1, \ldots, X_n$ are i.i.d. with $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$. Then as $n \to \infty$, 

$$
\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \text{Nor}(0, 1),
$$

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

Proof: Not in this class.
Remarks: (1) So 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 i.i.d.

(5) In fact, the CLT is actually a lot more general than the theorem presented here!
Example: Suppose $X_1, \ldots, X_{100} \overset{iid}{\sim} \text{Exp}(1/1000)$. Find $P(950 \leq \bar{X} \leq 1050)$.

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

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

$$E[\bar{X}] = E[X_i] = 1/\lambda \quad \text{and}$$

$$\text{Var}(\bar{X}) = \text{Var}(X_i)/n = 1/(n\lambda^2).$$

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

\[
P(950 \leq \bar{X} \leq 1050) = P\left( \frac{950 - E[\bar{X}]}{\sqrt{\text{Var}(\bar{X})}} \leq \frac{\bar{X} - E[\bar{X}]}{\sqrt{\text{Var}(\bar{X})}} \leq \frac{1050 - E[\bar{X}]}{\sqrt{\text{Var}(\bar{X})}} \right)
\]

\[
\approx P\left( \frac{950 - 1000}{100} \leq Z \leq \frac{1050 - 1000}{100} \right)
\]

\[
\approx P\left( -\frac{1}{2} \leq Z \leq \frac{1}{2} \right) = 2\Phi(1/2) - 1 = 0.383.
\]

Note: This problem can be solved exactly if we have access to the Excel Erlang cdf function \texttt{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 i.i.d. 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 $\approx 0.383$.

Notice that we didn’t care whether or not the data came from an exponential distrn. We just needed the mean and variance.
Normal Approximation to the Binomial\((n, p)\)

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 i.i.d. Bern\((p)\).

Then

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

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 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 cts. If you don’t want to use it, don’t worry too much.
Example: The Braves play 100 indep baseball games, each of which they have prob 0.8 of winning. What’s the prob that they win ≥ 84?

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

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

\[ \approx P\left(Z \geq \frac{83.5 - np}{\sqrt{npq}}\right) \quad \text{(CLT)} \]

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

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

$(X, Y)$ has the **Bivariate Normal Distrn** if it has pdf

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

where

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

$$ z_X(x) \equiv \frac{x - \mu_X}{\sigma_X} \quad \text{and} \quad z_Y(y) \equiv \frac{y - \mu_Y}{\sigma_Y}. $$
Pretty nasty joint pdf, eh?

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

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,

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

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,

\[
E[Y|X = 900] = \mu_Y + \rho(\sigma_Y / \sigma_X)(x - \mu_X) \\
= 2.3 + \rho(\sqrt{0.25/6400})(900 - 1300) = 0.8,
\]

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

Second,

\[
\text{Var}(Y|X = 900) = \sigma_Y^2(1 - \rho^2) = 0.16.
\]

Thus,

\[
Y|X = 900 \sim \text{Nor}(0.8, 0.16).
\]

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

**Turns Out:** The pdf and moments of the lognormal are

\[
f(x) = \frac{1}{x\sigma_Y \sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2_Y} [\ln(x) - \mu_Y]^2 \right\}, \quad x > 0,
\]

\[
E[X] = \exp\left\{ \mu_Y + \frac{\sigma^2_Y}{2} \right\}
\]

\[
\text{Var}(X) = \exp\{2\mu_Y + \sigma^2_Y\} \left( \exp\{\sigma^2_Y\} - 1 \right)
\]

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

\[
P(X \leq 1000) = P\left( Z \leq \frac{\ln(1000) - 10}{2} \right) = \Phi(-1.55) = 0.061.
\]
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** can calculate the inverse of the standard normal.

And you can use **RAND** to simulate a Unif(0,1) distribution. How would you simulate other RV’s? Well, this is the subject of another course, but here’s a remarkable example...
Simulating Normal RV’s

How would you generate a normal RV’s when conducting computer simulation experiments?

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

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

are i.i.d. $\text{Nor}(0,1)$.

Remarks: (1) Proof: Not here.

(2) Many other ways to generate $\text{Nor}(0,1)$’s, but this is the easiest.

(3) Cosine and Sine must be calculated in radians, not degrees.

(4) To get $X \sim \text{Nor}(\mu, \sigma^2)$, just take $X = \mu + \sigma Z$. 