You're watching a potentially endless sequence of Bernoulli trials with constant success probability $p$. Let $X = \# \text{ failures before } r^{th} \text{ success}$, where $r$ is an integer $\geq 1$. You can show that $X$ follows the Negative Binomial distribution.

Its PDF is:

$$f(x | r, p) = \binom{r+x-1}{x} p^r (1-p)^x$$

with parameters $(r, p)$, $(0 < p < 1)$, $x \in \{0, 1, 2, \ldots \}$.

The name comes from the fact that, when you watch a sequence of Bernoulli trials with an unknown constant success probability $p$, there are two different ways to unfold.
(254)

Estimate $p$; decide ahead of time to sample $n$ successes/failures trials and record the (random) # $S$ of successes you see (from which a reasonable estimate would be $\hat{p} = \frac{S}{n}$).

or decide ahead of time that you're going to sample until you've seen $S$ known (constant) successes & record the (random) # of trials $N$ needed to accumulate that many successes (from which a reasonable estimate would be $\hat{p}_{NB} = \frac{S}{N}$ (Negative Binomial)).
Set \( r = 1 \) and record the \( X \) number of failures until the first success: \( X \) is said to follow the

Geometric (\( p \)) distribution, with

\[
P(X = x | p) = p(1-p)^x \quad x \in \{0, 1, \ldots\}
\]

(parameter \( p \))

\( \{X_1, \ldots, X_r\} \) IID Geometric(\( p \))

\[
\sum_{i=1}^{r} X_i \sim \text{Negative Binomial}(r, p)
\]

This is a direct analogue to the

Bernoulli / Binomial story: \( \{X_1, \ldots, X_r\} \) IID

Bernoulli (\( p \)) \( \Rightarrow \sum_{i=1}^{r} X_i \sim \text{Binomial}(r, p) \).
$X \sim \text{Negative Binomial } (r, p)$

$\psi_X(t) = \left[ \frac{p}{1 - (1-p)e^t} \right]^r \quad \text{for } t < \log \left( \frac{1}{1-p} \right)$

From which $E(X) = \frac{r(1-p)}{p}$, $\text{Var}(X) = \frac{r(1-p)}{p^2}$

Consequence $X \sim \text{Geometric } (p) \Rightarrow$

$\{ k \geq 0 \text{ both non-negative integers} \}
\quad P(X = k + t | X \geq k) = P(X = t)$

This is called the memoryless property of the Geometric distribution, and it turns out that this is the only
discrete distribution with this property.

$X$ = # failures until first success = 5

$Y$ = # failures, starting at trial $(k+1)$ until next success

~4 here

Then $Y$ has the same dist. as $X$ and is independent of what happened on the first $k$ trials, i.e., "the process has no memory".

**Important Case 2: Continuous Distributions**

Normal $\mathcal{N}(\mu, \sigma^2)$ mean $\mu$, variance $\sigma^2 < \infty$

PDF

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$
The Normal distribution is the single most important distribution in all of probability and statistics, mainly for 2 reasons:

1. Many observable random processes have distribution shapes that are close to the bell curve (Normal PDF), and
2. The Central Limit Theorem (CLT), which we'll examine soon.

Properties of the Normal Distribution:

- \( \Sigma \sim \text{Normal} (\mu, \sigma^2) \)
- \( \mathbb{E}(\Sigma) = \mu \)
- \( \text{Var}(\Sigma) = \sigma^2 \)
- \( \text{SD}(\Sigma) = \sigma \)
- \( \Psi_{\Sigma}(t) = \exp(\mu t + \frac{\sigma^2 t^2}{2}) \)
- Center of symmetry = mean = median = mode = \( \mu \)
Consequences

1. $X \sim \text{Normal}(\mu, \sigma^2)$,

$Y = aX + b$, ($b \neq 0$) fixed constant $\rightarrow$

$Y \sim \text{Normal}(a\mu + b, a^2\sigma^2)$.

In other words, Normality is preserved under linear transformation $\text{Def.}$

The Normal dist. with mean $\mu = 0$ and SD $\sigma = 1$ is called the \textit{standard normal dist.}.

The PDF of $\text{Normal}(0, 1)$ is

$\phi_X(x) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{x^2}{2} \right)$ and its lower-case $\phi$ (phi)

CDF is $F_X(x) = \int_{-\infty}^{x} \phi_X(t) \, dt$
It turns out that \( e^{-\frac{x^2}{2}} \) has no anti-derivative in closed form, so \( \Phi(x) \) cannot be summarized in a formula; instead it's approximated by numerical integration (see p. 561 in DS).

Consequently, because the Normal PDF (for all \( x \in \mathbb{R} \)) is symmetric, \( \Phi(-x) = 1 - \Phi(x) \) and \( \Phi^{-1}(p) = - \Phi^{-1}(1-p) \) (for all \( 0 < p < 1 \)).

3. \( \xi \sim \text{Normal} (\mu, \sigma^2) \Rightarrow \frac{\xi - \mu}{\sigma} \sim \mathcal{N}(0,1) \), so that \( F_{\frac{\xi}{\sigma}}(x) = \Phi\left( \frac{x - \mu}{\sigma} \right) \), and \( F_{\frac{\xi}{\sigma}}^{-1}(p) = \mu + \sigma \Phi^{-1}(p) \).
Empirical Rule

Part 1: Start at the mean, $m$.

Part 2: 68% of the probability in the interval $(m \pm 1 \sigma)$. Ditto 2SDs.

Part 2: Ditto 2SDs.

Part 3: Ditto 3SDs either way: $(m \pm 3 \sigma)$ capture almost all 99.7% of the probability. This rule is exact for all Normal dists & is a surprisingly
good approximation for many other distributions, this permits an easy trick that's helpful in computing Normal probabilities. Example: Random sample of $n = 103$ immature monarch butterflies, and you measure their wing lengths:

- $y_1 = 4.1$
- $y_2 = 3.3$
- $\vdots$
- $y_{103} = 4.7$

**mean $\bar{y} = 3.96$ cm**

**$SD \ s = 0.29$ cm**

**histogram of wing length**

3.5 3.96 cm

**Q: About what % of the sampled butterflies had wing length $\leq 3.5$ cm?**

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Sample mean

Sample SD
For details, see section on continuity correction.

- \( Z = \frac{t - \mu}{\sigma} = \frac{2 - 1.25}{0.25} = 3.5 \) for random variables

- \( z = 3.5 \) - 3.96

- From \( Z = 0.85 \) to 1.58

- Unit units

- Unit units

- \( Z = 1.96 \) - 0.05

- \( \mu = 3.5 \) - 3.96

- \( \sigma = 0.25 \) - 0.52

- 0.25
4. \( X_1, \ldots, X_k \) independent

\[
X_i \sim \text{Normal}(\mu_i, \sigma_i^2)
\]

\[
\sum_{i=1}^{k} X_i \sim \text{Normal} \left( \sum_{i=1}^{k} \mu_i, \sum_{i=1}^{k} \sigma_i^2 \right)
\]

This is the additive property. This is why Normal dists are indexed by variance rather than SD.

**Notation**

\[
\text{Normal}(\mu, \sigma^2) \equiv N(\mu, \sigma^2)
\]

**Example**

Population of women: height follows \( N(65.0\text{in}, \sigma^2 = 3.2\text{in}^2) \) dist.

\( \sigma = 3.2\text{in} \)

Pop. of adult U.S. men: height follows \( N(\mu = 69.5\text{in}, \sigma^2 = 3.3\text{in}^2) \) dist.
1 woman chosen at random, height \( w \); 1 man chosen at random (independently), height \( M \); \( P( \text{woman taller than man} ) \)

Define \( J = W - M \)

\[ P( W > M ) = P( J > 0 ) \]

By consequence \( 4 \), \( D \sim N(65 - 69.5 = -4.5, \text{in}^2) \)

\[ 3.2^2 + 3.3^2 = 21.1 \text{ in}^2 \]

\[ \text{SD} \sqrt{21.1 \text{ in}^2} = 4.6 \text{ in} \]

Convert to \( Z \):

\[ 0 - (-4.5) = 4.5 \]

\[ z \approx 0.98 \]

From table, \( 0.8365 \)

\[ 1 - 0.8365 = 0.1635 \]

So \( P( W > M ) = 16\% \)
Def \( rv X_1, \ldots, X_n \rightarrow \) sample mean

If \( (X_1, \ldots, X_n) \) is \( \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \)

Consequently, \[ \{ X_i \sim N(\mu, \sigma^2) \} \]

\[ (i = 1, \ldots, n) \]

\[ + \bar{X}_n \sim N(\mu, \frac{\sigma^2}{n}) \]

So \( SD(\bar{X}_n) = \frac{\sigma}{\sqrt{n}} \)

Because \( E(\bar{X}_n) = \mu \), \( \bar{X}_n \) is an \textbf{unbiased estimator of} \( \mu \)

In frequentist statistics, the standard deviation (SD) of an estimator \( \hat{\theta} \) of a parameter \( \theta \) is called the \textbf{standard error SE(\( \hat{\theta} \))} of \( \hat{\theta} \).
So if you use $\bar{X}_n$ as an estimate of $\mu$, $SE(\bar{X}_n) = \frac{\sigma}{\sqrt{n}} \to 0 \text{ as } n \to \infty$

As $n \uparrow$, $\bar{X}_n$ gets better as an estimate of $\mu$, at a $\sqrt{n}$ rate; this is called the **square root law**.

Unfortunately, this means that to cut the $SE(\bar{X}_n)$ in half, you have to **quadruple** the sample size.
log normal (This distribution is mis-named!
Distribution it should be called the
Exponential Normal distribution, but
we're stuck with a bad name.)

\[ x > 0 \]

If \( Y = \log(X) \sim N(\mu, \sigma^2) \), people
say that \( X \sim \text{Log Normal} (\mu, \sigma^2) \).

\[
\begin{align*}
\text{PDF of } Y & \rightarrow \text{PDF of } Z = e^Y \\
\mu & \rightarrow 80 \\
X & \sim \text{Log Normal} (\mu, \sigma^2) \\
Y & = \log(X) \sim N(\mu, \sigma^2)
\end{align*}
\]
The MGF of $X$ is $\phi_X(t) = \exp(\mu t + \frac{1}{2} \sigma^2 t^2)$

But by definition

$\phi_X(t) = E(e^{tX}) = E(e^{\log X})$

$= E(X^t)$, so we can

read the moments of $X$

directly from the MGF

of $X$

$\sqrt{\text{Var}(X)} = \phi_X(2) - \left(\phi_X(1)\right)^2$

$= \exp(2\mu + \sigma^2) \left(e^{\sigma^2} - 1\right)$.

Famous pricing stock options, continued

example

1 share of a stock, current price $S_0$. Heroic assumption: price
u time units in the future will be.

$$S_u = S_0 e^{z_u}, \quad z_u \sim N(\mu_u, \sigma^2_u).$$

Can write $S_0 e^{z_u} = e^{z_u + \log(S_0)}$.

Now

$$[z_u + \log(S_0)] \sim N(\mu_u + \log(S_0), \sigma^2_u),$$

so $S_u \sim \log\text{Normal}(\mu_u + \log(S_0), \sigma^2_u)$.

Consider a single time horizon $u$;

[hermetic assumption revised]

$$S_u = S_0 e^{[\mu_u + (\sigma_u/2)],}$$

$z_1 \sim N(0, 1)$

we need to price the option to buy 1 share of this stock for price $g$ at time $u$. 
Use risk-neutral pricing as in the previous discussion: force present value $E(S_u) = S_0$. Let time scale of $u$ be in years; let risk-free (continuous compounding) interest rate be $r/\text{year}$. Then present value of $E(S_u)$ is $e^{-ru}E(S_u)$.

But by heroic assumption, $S_u$ is lognormal.

Recall $E(S_u) = S_0 \exp \left( \mu u + \frac{\sigma^2 u}{2} \right)$.

The result is $\left( \mu = r - \frac{\sigma^2}{2} \right) e^{-ru} S_0 \exp \left( \mu u + \frac{\sigma^2 u}{2} \right)$ for risk-neutral pricing.
Value of option at time $t$ will be $h(S_t)$, where

$$h(S_t) = \begin{cases} S - 3 & \text{if } S > 3 \\ 0 & \text{else} \end{cases}$$

with $\mu = r - \frac{\sigma^2}{2}$, $h(S_t) > 0$ iff

$$\frac{1}{2} > \log \left( \frac{S_t}{S_0} \right) - (r - \frac{\sigma^2}{2}) t \quad \frac{\sigma \sqrt{t}}{2} \mu = c$$

Now a nasty integral gives: risk-neutral price of option is the present value of $E[h(S_t)]$, which is

$$E[h(S_t)] = e^{-r_t} \int_{\delta\text{e}}^{\infty} (r - \frac{\sigma^2}{2}) t + \sigma \sqrt{t} u - 2 \delta \mu.$$

Careful calculation reveals the (famous) formula

$$\sqrt{2\pi} \int_{-\infty}^{\infty} e^{-x^2/2} dx = 1.$$
The risk-neutral price of an option is given by the Black-Scholes formula:

\[ \frac{\partial F}{\partial t} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 F}{\partial S^2} + rS \frac{\partial F}{\partial S} - rF = 0 \]

where

\[ e^{-rT} E \left( \frac{S_T - K}{S_0} I(\Delta > 0) \right) \]

\[ (\sigma^2 + \rho \sigma \beta) \]
PDF \( f_X(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} I(x>0) \)

d is called a shape parameter in the \( \Gamma(\alpha, \beta) \) family because it governs things like skewness of the dist. \( \beta \) is related to the scale of the distribution, which measures how spread out the dist. is. \( \Gamma(x) \) is the Gamma function, invented to deal with integrals of functions like above: \[ \Gamma(x) = \int_0^\infty x^{x-1} e^{-x} dx \]

\( \Gamma(x) \) has no anti-derivative in closed form.
\(\Gamma(x)\) turns out to be a continuous generalization of the factorial function, because \((n \text{ positive integer}) \rightarrow \Gamma(n) = (n-1)!

\(\Gamma(x) \rightarrow \infty\) really quickly as \(x \rightarrow \infty\), so it's better to evaluate the Gamma PDF on the log scale and then exponentiate:

\[
\int e^{-x} x^{\beta - 1} \, dx = \exp \left( \int \left[ \beta \log(x) - \log(\Gamma(\beta)) + (\beta - 1) \log(x) - \beta x \right] \, dx \right)
\]

Another way to take \(\Gamma(x)\) is with a Stirling's approximation:

\[
\Gamma(x) \approx \sqrt{2\pi} x^{x - \frac{1}{2}} e^{-x}
\]

for large \(x\).
Exponential distribution

But this is just our old friend

Special care

Alternative

\( E(X) = \frac{1}{\beta} \)

with

\( \lambda = 1 \) Hence \( \lambda = \frac{1}{\beta} \)

for \( t > 0 \)

\( \beta \)

so \( \text{CDF} \)

\( F(x) = \left(1 - \frac{x}{\beta}\right)^{-\lambda} = (x - 0) \) for \( t > 0 \)

\( \frac{\lambda^\lambda}{(\lambda - 1)^{\lambda - 1}} \)

\( x \sim F(x) \)

\( \gamma(t) = \left(1 - \frac{t}{\beta}\right)^{-\lambda} \)

\( \frac{\lambda^\lambda}{(\lambda - 1)^{\lambda - 1}} \) for \( t > 0 \)

\( \frac{x}{\beta} \)

\( \frac{1}{\beta} \)

\( \frac{1}{\beta} \) for \( t > 0 \)

\( \frac{1}{\beta} \) for \( t > 0 \)

\( \frac{1}{\beta} \) for \( t > 0 \)
$X \sim \text{Exponential}\left(\beta\right)$

$T = \frac{\beta}{X}$

$V(x) = \frac{1}{\beta^2}$

Notice that the Exponential distribution looks like a gamma distribution, but the parameter $\beta$ is related to the distribution of the waiting time to the first event.

**Never Forget**

[Text on diagram]

Not arrived yet (events) occur at a rate of $\lambda = \frac{1}{2}$.
The $Y_i$ are called the inter-arrival times. Then it turns out that $Y_i \sim \text{Exponential}(\beta)$.

The Exponential dist. is also related to the Geometric dist., in that they both have a memory less property. Theorem

$X \sim \text{Exponential}(\beta); \quad t > 0, \quad h > 0$

$\Rightarrow P(X \geq t + h \mid X \geq t) = P(X \geq h)$

Example\[ X = \text{time until a manufactured product fails} \]

$F_X(x) = P(X \leq x) = 1 - F_X(x) = P(X > x)$

$= P(\text{"system surviving" at least to time } x)$
For this reason, \( 1 - F_\mathcal{X}(x) \) is called the survival function \( S_\mathcal{X}(x) = 1 - F_\mathcal{X}(x) \) in medicine and the reliability function \( R_\mathcal{X}(x) = 1 - F_\mathcal{X}(x) \) in engineering.

Earlier we showed that \( F_\mathcal{X}(x) = 1 - e^{-\beta x} \) for \( \mathcal{X} \sim \text{Exponential (}\beta) \) for \( x > 0 \).

So \( S_\mathcal{X}(x) = R_\mathcal{X}(x) = e^{-\beta x} \) for this dist.

The instantaneous failure or hazard rate function is defined to be \( h_\mathcal{X}(x) = \frac{f_\mathcal{X}(x)}{S_\mathcal{X}(x)} = \frac{f_\mathcal{X}(x)}{1 - F_\mathcal{X}(x)} \) for small \( \ell \) for interval \((x, x + \ell)\) to time \( x \)
Notice that if $X \sim \text{Exponential} (\beta)$, then

$$H_X(x) = \frac{\beta e^{-\beta x}}{e^{-\beta x}} = \beta \left( \frac{\text{Constant in}}{x} \right)$$

The Exponential is the only failure rate distribution with constant hazard. Returning to the earlier result that $X \sim \text{Exponential} (\beta)$,

$$P(X \geq t+h \mid X \geq t) = P(X \geq h),$$

for all $t > 0$, $h > 0$. This says that if the product has survived to time $t$, the chance it will survive to time $(t+h)$ is the same as the original chance of surviving from time $0$ to time $h$, i.e., "the system doesn't remember how long it's survived".
Consequences

1. $X_1 \sim \text{Exponential} (\beta)$
   
   $(x^i, \ldots, X_n)$,

   then

   $Y_1 = \min (X_1, \ldots, X_n) \sim \text{Exponential} (\eta \beta)$.

Beta

$x, \beta > 0 \rightarrow X \sim \text{Beta} (\alpha, \beta) \leftrightarrow$

Distribution

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha) \Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \cdot$$

$I(0 < x < 1)$

The name comes from the normalizing constant: the function $x^{\alpha-1} (1-x)^{\beta-1}$ has no closed-form anti-derivative, so people just make

$$\textbf{Definition} \quad \text{For all } \alpha > 0, \beta > 0 \Rightarrow B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx$$

beta function
Can show that $B(\alpha, \beta) = \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha + \beta)}$. (282)

$(\alpha, \beta)$ jointly control the shape of the Beta$(\alpha, \beta)$ dist.

$X \sim \text{Beta}(\alpha, \beta)$

$\mathbb{E}(X) = \frac{\alpha}{\alpha + \beta}$

$\text{Var}(X) = \left( \frac{\alpha}{\alpha + \beta} \right) \left( \frac{\beta}{\alpha + \beta} \right) \left( \frac{1}{\alpha + \beta + 1} \right)$

Example (Castaneda v. Partida continued)