*There will be an extra lecture next Wednesday evening through Webcast -> more details soon*

**Definition**

$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$

**Consequences continued**

5) $\bar{X}_i \xrightarrow{iid} \mathcal{N}(\mu, \sigma^2)$ ($i = 1, \ldots, n$) (All of them)

$\Rightarrow \bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ (Each of them)

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

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

**Definition**

In frequentist statistics, the standard deviation (SD) of an estimator $\hat{\theta}$ of a parameter $\theta$ is called the standard error $SE(\hat{\theta})$ of $\hat{\theta}$

So if you use $\bar{X}_n$ as an estimate of $\mu$, $SE(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$

$SE(\hat{\theta}) = \frac{\sigma}{\sqrt{n}}$

**PDF of $\bar{X}_n$, $\mathcal{N}$**

This is the basis of most frequentist statistical inference

As $n \to \infty$, $\bar{X}_n$ gets better as an estimate of $\mu$, at $\frac{1}{\sqrt{n}}$ rate. This is called the Square Root Law.
This means that to cut the $SE(\bar{x}_n)$ in half, you have to quadruple the sample size.

**Lognormal Distribution should be exponential normal**

**Def:** If $\bar{x}$ is a mean and $\bar{x} = \log(\bar{x}) \sim N(m, \sigma^2)$, then $\bar{x} \sim \text{Lognormal}(m, \sigma^2)$

$\bar{x} \sim \text{Lognormal}(m, \sigma^2)$

$\bar{x} = \log(\bar{x}) \sim N(m, \sigma^2)$

Can get MGF of $\bar{x}$ from MGF of $\bar{z}$

**MGF of $\bar{z}$:** $\phi_{\bar{z}}(t) = \exp(m + \frac{1}{2} \sigma^2 t^2)$

But by def: $\phi_{\bar{z}}(t) = E(e^{t\bar{z}}) = E(e^{t\log\bar{x}}) = E(\bar{x}^t)$

So we can read the moments of $\bar{x}$ directly from the MGF of $\bar{z}$

$E(\bar{x}) = \phi_{\bar{z}}(1) = \exp(m + \frac{1}{2} \sigma^2)$

$V(\bar{x}) = \phi_{\bar{z}}(2) - [\phi_{\bar{z}}(1)]^2 = \exp(2m + \sigma^2)[e^{\sigma^2} - 1]$
Gamma distribution

$(\alpha, \beta > 0)$ $X$ has the Gamma dist. with parameter $(\alpha, \beta)$, written $X \sim \Gamma(\alpha, \beta)$ or $X \sim \text{Gamma}(\alpha, \beta)$

$\Rightarrow X$ continues on $(0, \infty)$ with...

PDF $f_X(x) \sim \Gamma(\alpha, \beta)$

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

$\alpha$ 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 distribution is

$\Gamma(\alpha)$ is the Gamma function, inverted to deal with integrals of functions like $\circ$ above:

\[
\Gamma(\alpha) = \int_0^\infty x^{\alpha - 1} e^{-x} \, dx
\]

\[
\uparrow
\]

has no anti-derivative in closed form

$\Gamma(\alpha)$ turns out to be a continuous generalization of the factorial function, because (in positive integer)

$\Rightarrow \Gamma(n) = (n-1)!$
\[
\Gamma(x) \to \infty \text{ really quickly as } x \to \infty, \text{ so it's better to evaluate the Gamma PDF on the log scale and then exponentiate:}
\]
\[
\frac{\beta^a}{\Gamma(a)} x^{a-1} e^{-\beta x} = \exp \left[ \alpha \ln(\beta) - \ln \Gamma(a) + (\alpha - 1) \ln(x) - \beta x \right]
\]

Another key to tame \( \Gamma(x) \) is with a Stirling's approximation

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

so that \( \ln \Gamma(x) \approx \frac{1}{2} \ln(2\pi) + (x - \frac{1}{2}) \ln x - x \)

\[X \sim \Gamma(\alpha, \beta)\]

\[\psi_X(t) = (1 - \frac{t}{\beta})^{-\alpha} \text{ for } t < 2\beta\]

so \( E(X) = \frac{\alpha}{\beta} \) and \( V(X) = \frac{\alpha}{\beta^2} \) so \( \sigma(X) = \frac{\sqrt{\alpha}}{\beta} \)

Alternate expression

\[\psi_X(t) = \left(\frac{\beta}{\beta - t}\right)^\alpha \text{ for } t < 2\beta\]

Special case of \( \Gamma(\alpha, \beta) \)

With \( \alpha = 1 \) the PDF is \( \rho_X(x | 1/\beta) = \beta e^{-\beta x} I(x > 0) \)

This is just the exponential distribution.
\( X \sim \text{Exponential } (\beta) \)

\[
\Psi_X(t) = \frac{\beta}{\beta + 1 + 2\beta} \\
E(X) = \frac{1}{\beta} \\
V(X) = \frac{1}{\beta^2} \\
S_D(X) = \frac{1}{\beta} \\
\int f_X(x) = 1 \\
\beta \\
1 \\
x
\]

**Theorem:** Suppose that arrivals (events) occur according to a Poisson process with rate \( \beta \) per unit time. and define

\[
Z_1 = Z_1 - 0 \\
Z_2 = Z_2 - Z_1 \\
\vdots \\
Z_k = Z_k - Z_{k-1} \quad \text{for } k = 2, 3, \ldots
\]

Set \( Z_k = \) time until \( k \)-th arrival \( k = 1, 2, \ldots \)

The \( Z_i \) are the inter-arrival times.

\( Z_i \) \( \sim \) \text{Exponential } (\beta) \quad \text{as a result.}

The exponential dist. is also related to the Geometric dist. in that they both have a memoryless property.

**Theorem:** \( X \sim \text{Exponential } (\beta) \); \( t > 0; \lambda > 0 \)

\[
P(X > t + h \mid X > t) = P(X > h)
\]

**Example:** \( X = \) time from initial use until a manufactured product fails (e.g. light bulb)
\[ F_\bar{X}(x) = P(\bar{X} \leq x) \]
\[ 1 - F_\bar{X}(x) = P(\bar{X} > x) \]
\[ = P(\"system survives\" \text{ at least to time } x) \]

For this reason, \( 1 - F_\bar{X}(x) \) is the survival function \( S_\bar{X}(x) = 1 - F_\bar{X}(x) \) in medicine and the reliability function \( R_\bar{X}(x) = 1 - F_\bar{X}(x) \) in engineering.

For \( \bar{X} \sim \text{Exponential}(\beta) \rightarrow F_\bar{X}(x) = 1 - e^{-\beta x} \text{ for } x \geq 0 \)

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

The instantaneous failure rate or hazard rate function is defined to be
\[
H_\bar{X}(x) = \frac{f_\bar{X}(x)}{S_\bar{X}(x)} = \frac{f_\bar{X}(x)}{R_\bar{X}(x)}
\]

This gives \( P(\text{failure in interval } (x, x+\epsilon) \mid \text{survival to time } x) \) for small \( \epsilon \)

Notice that if \( \bar{X} \sim \text{Exponential}(\beta) \) then \( H_\bar{X}(x) = \frac{\beta e^{-\beta x}}{e^{-\beta x}} = \beta \) (constant in \( x \))

The exponential is the only failure rate distribution with constant hazards.

Returning to the earlier result that \( \bar{X} \sim \text{Exponential}(\beta) \)

\[ \to P(\bar{X} \geq t + h \mid \bar{X} \geq t) = P(\bar{X} \geq h) \]

For all \( t \geq 0, h \geq 0 \)
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.

"the system doesn't remember how long it survived"

(This often makes the Exponential unrealistic in practice)

Consequences

1) $\Xi_i \sim \text{Exponential} (\lambda_i) (i = 1, \ldots, n)$ then

\[ \Xi = \min (\Xi_1, \ldots, \Xi_n) \sim \text{Exponential} (\lambda) \]

Beta Distribution

\[ a, \beta > 0 \]

\[ \Xi \sim \text{Beta} (a, \beta) \leftrightarrow f_\Xi (x) = \frac{\Gamma (a+\beta)}{\Gamma (a) \Gamma (\beta)} x^{a-1} (1-x)^{\beta-1} \]

\[ I (0 < x < 1) \]

The name comes from the normalizing constant; the function $x^{a-1} (1-x)^{\beta-1}$ has no closed-form anti-derivative, so people just made a definition...

Def: For all $a > 0, \beta > 0$

\[ B (a, \beta) = \int_0^1 x^{a-1} (1-x)^{\beta-1} \, dx \]

\[ \uparrow \]

beta function
Can show that \( B(d_1, \beta) = \frac{\Gamma(d_1) \Gamma(\beta)}{\Gamma(d_1 + \beta)} \)

\((d_1, \beta)\) jointly control the shape of the \(B (d_1, \beta)\) dist.

\(\bar{X} \sim B (d_1, \beta)\)

\[ f_{\bar{X}}(+) = 1 + \sum_{k=1}^{\infty} \frac{1}{k!} \prod_{r=0}^{k-1} \frac{d+r}{d+\beta+r} \cdot \frac{t^k}{k!} \]

\[ E(\bar{X}) = \frac{d}{d+\beta} \]

\[ V(\bar{X}) = \left( \frac{d}{d+\beta} \right) \left( \frac{\beta}{d+\beta} \right) \left( \frac{1}{d+\beta+1} \right) \]

Multinomial Distributions (back to discrete)

You're contemplating a population that contains elements of \( k \geq 2 \) types

(e.g., Democrat, Republican, Libertarian, Independent, Green)

Suppose the proportion of elements of type \( i \) is \( \pi_i \), \( \sum_{i=1}^{k} \pi_i = 1 \)

with \( \bar{p} = (p_1, \ldots, p_k) \)

You take an \( 110 \) sample of size \( n \) from this pop.

\( \bar{X}_i = \# \) elements of type \( i \) in your sample

\[ \sum_{i=1}^{k} \bar{X}_i = n \]

Can show that the vector \( \bar{X} = (\bar{X}_1, \ldots, \bar{X}_k) \) has p.m.f.

\[ f_{\bar{X} \mid \bar{p} = \bar{X}} (\bar{X} \mid \bar{p}) = \sum (x_1, \ldots, x_k) p_1^{x_1} \ldots p_k^{x_k} \]

\[ \begin{cases} 
\pi_1^{x_1} \ldots \pi_k^{x_k} \text{ if } \sum_{i=1}^{k} x_i = n \\
0 \text{ else} 
\end{cases} \]
where \( \binom{n}{x_1, \ldots, x_k} = \frac{n!}{x_1! x_2! \cdots x_k!} \) is the multinomial coefficient.

This is called the Multinomial \((n, \mathbf{p})\) distribution.

\[
\begin{align*}
E(\bar{X}_i) &= np_i \\
\text{Var}(\bar{X}_i) &= np_i(1-p_i) \\
C(\bar{X}_i, \bar{X}_j) &= -np_i p_j
\end{align*}
\]

Negatively correlated because \( \sum_{i=1}^{k} \bar{X}_i = n \).

**Bivariate Normal Dist.**

Can build a 2-dimensional (bivariate) version of the normal dist. as follows:

\[
\begin{align*}
&\begin{array}{c}
\begin{tikzpicture}
\fill[white,opacity=0.3] (0,0) circle (2 cm);
\fill[white,opacity=0.3] (-1.5,0) -- (1.5,0) -- (0,1) -- cycle;
\draw[->] (-2,0) -- (2,0) node[right] {\(x_1\)};
\draw[->] (0,-2) -- (0,2) node[above] {\(x_2\)};
\draw[domain=-1:1] plot ({\x},{\x^2/4}) node[right] {\(M_2\)};
\end{tikzpicture}
\end{array}
\end{align*}
\]

\[
\begin{align*}
&Z_1, Z_2 \sim N(0,1) \\
&\text{specify 5 parameters} \\
&1) \ -\infty < \mu_1 < \infty \\
&2) \ -\infty < \mu_2 < \infty \\
&3) \ 0 < \sigma_1 < \infty \\
&4) \ 0 < \sigma_2 < \infty \\
&5) \ -1 < \rho < 1
\end{align*}
\]

Now build \((\bar{X}_1, \bar{X}_2)\) with the transformation

\[
\begin{align*}
\bar{X}_1 &= \mu_1 + \sigma_1 Z_1 \\
\bar{X}_2 &= \sigma_2 \left[ \rho Z_1 + \sqrt{1-\rho^2} Z_2 \right] + \mu_2
\end{align*}
\]

The joint PDF of \(\bar{X} = (\bar{X}_1, \bar{X}_2)\) is then

\[
\begin{align*}
f_{\bar{X}_1, \bar{X}_2}(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \exp \left\{ \frac{1}{2(1-\rho^2)} \left[ \frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1 \sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \right] \right\}
\end{align*}
\]
This is the Bivariate Normal \((M_1, M_2, \sigma_1, \sigma_2, \rho)\) dist.

Easy to show that \(E(X_1) = M_1\), \(E(X_2) = M_2\)

\(\nu(X_1) = \sigma_1^2\), \(\nu(X_2) = \sigma_2^2\), \(\rho(X_1, X_2) = \rho\)

**Consequences of this def:**

1) \((X_1, X_2) \sim \text{Bivariate Normal} \Rightarrow \begin{pmatrix} X_1, X_2 \end{pmatrix} \sim \begin{pmatrix} \text{independent} \end{pmatrix} \Rightarrow \begin{pmatrix} \text{uncorrelated} \end{pmatrix}\)

We already knew the \(\Rightarrow\) direction in general. What's new here is that correlation 0 implies independence if

\((X_1, X_2) \sim \text{Bivariate Normal}\)

2) \((X_1, X_2) \sim \text{Bivariate Normal} (M_1, M_2, \sigma_1, \sigma_2, \rho) \Rightarrow\)

conditional distribution of \(X_2\) given that \(X_1 = X_{1*}\) is

(univariate) normal with mean \(E(X_2 | X_1) = M_2 + \frac{\rho \sigma_2}{\sigma_1} (X_1 - M_1)\)

and variance \(\nu(X_2 | X_1) = (1 - \rho^2) \sigma_2^2\)

Galton  
result 2 says that if \((X_1, X_2)\) are Bivariate Normal then the conditional distributions of \(X_2\) given \(X = X_{1*}\) in all of the vertical strips are also normal.

And the means of all these normal distributions in the vertical strips are connected together by Galton's regression line

\[ X_2^* = M_2 + \frac{\rho \sigma_2}{\sigma_1} (X_1 - M_1) \]

\[ X_2^* = \beta_0 + \beta_1 X_1 \]

This line has slope \(\beta_1 = \rho \frac{\sigma_2}{\sigma_1}\) and "y"-intercept \(\beta_0 = M_2 - \beta_1 M_1\)
Moreover, we can now quantify an earlier insight:

\[ \hat{x}_2 \] predict \( \hat{x}_2 = \mu_2 = E(\bar{x}_2 | x_1) \]

(root mean squared error) (RMSE) of this prediction is

\[ \sqrt{V(\bar{x}_2)} = \sigma_2 \]

use \( x_1 \) to predict \( \bar{x}_2 \)

\[ \text{predict } (\bar{x}_2)_{\text{use}} = E(\bar{x}_2 | \bar{x}_1 = x_1) \]

\[ = \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x_1 - \mu_1) \]

RMSE of this prediction is

\[ \sqrt{V(\bar{x}_2 | x_1)} = \sigma_2 \sqrt{1 - \rho^2} \]

Since \(-1 \leq \rho \leq 1\), \( \sigma_2 \sqrt{1 - \rho^2} \geq \sigma_2 \) with equality only when \( \rho = 0 \).

3) \((\bar{x}_1, \bar{x}_2) \sim \text{Bivariate Normal}(\mu, \mu_2, \sigma_1, \sigma_2, \rho)\),

\[ Y = a_1 \bar{x}_1 + a_2 \bar{x}_2 + b, \quad (a_1, a_2, b) \text{ arbitrary constants} \]

\[ \Rightarrow \bar{Y} \sim N(\mu + a_2 \mu_2 + b, a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + 2a_1 a_2 \rho \sigma_1 \sigma_2) \]

Large random samples

You draw an IID random sample \( \bar{x}_1, \ldots, \bar{x}_n \) from a population, with the goal of estimating the population mean \( \mu = E(\bar{x}_1) \).

We've already seen that from a root mean squared error point of view, the sample mean \( \bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} \bar{x}_i \) is the best you can do (in the absence of prior information).

It would be nice if \( \bar{x}_n \) approached the right answer \( \mu \) as \( n \) increases; how to quantify that idea?
Two inequalities that help

Markov Inequality

Suppose $\bar{X}$ is a non-negative rv, ie $P(\bar{X} \geq 0) = 1$

Then for all real $t > 0$, $P(\bar{X} \geq t) \leq \frac{E(\bar{X})}{t}$ says that if $E(\bar{X})$ is fixed, you can't move more and more probability out into the right tail beyond a certain point.

Ex: $E(\bar{X}) = 1$, $\bar{X}$ non-negative $\Rightarrow P(\bar{X} \geq 100) \leq \frac{1}{100}$

The inequality is sharp, meaning that the upper bound $\frac{E(\bar{X})}{t}$ on $P(\bar{X} \geq t)$ is attainable but most of the time it's a crude upper bound.

Ex: $E(\bar{X}) = 1$, $\bar{X}$ non-negative $\Rightarrow$ put probability 0.99 on $\bar{X} = 0$ and probability 0.01 on $\bar{X} = 100$. 