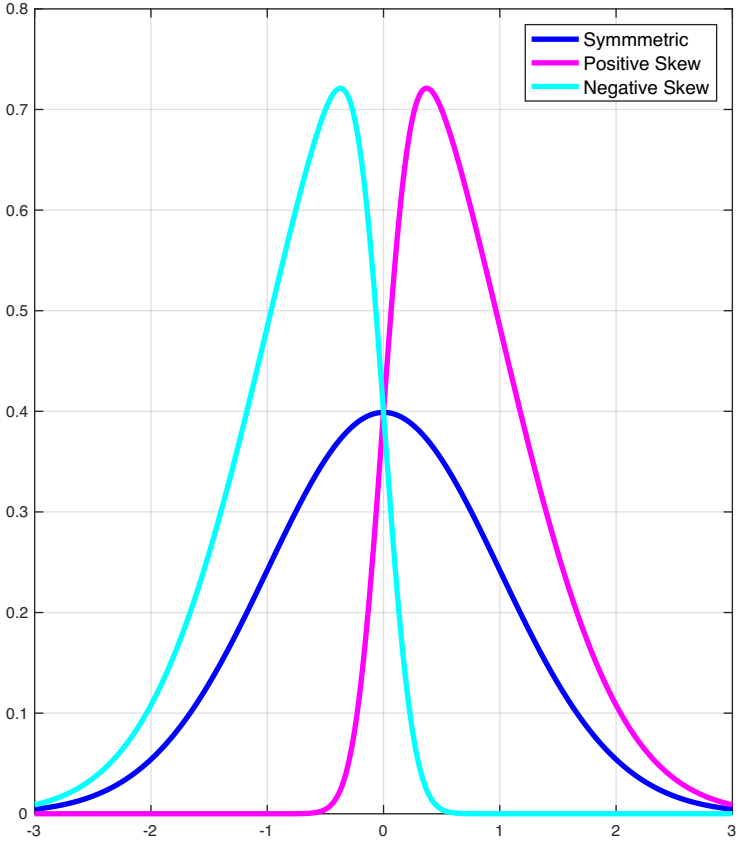
Random Variables & Sampling

Properties of a random variable

- **Mean Absolute Deviation**
  - Also measures dispersion, but does not magnify large deviations from the mean by squaring.
  - General formula: $\mathbb{E}\left[|X - \mathbb{E}[X]|\right]$
  - Discrete random variable: $\sum_{i=1}^{n} p_i |X_i - \mu_X|$
  - Continuous random variable: $\int |x - \mu_X| f(x) dx$
- **Skewness**
  - Measures how much a distribution deviates from symmetry.
  - General formula: $\mathbb{E}\left[(X - \mathbb{E}[X])^3\right] / \sigma_X^3$
  - Unit free
  - Symmetric distribution: skew $= 0$
  - skew > 0: right tail of the distribution is longer than the left tail
  - skew < 0: left tail of the distribution is longer than the right tail

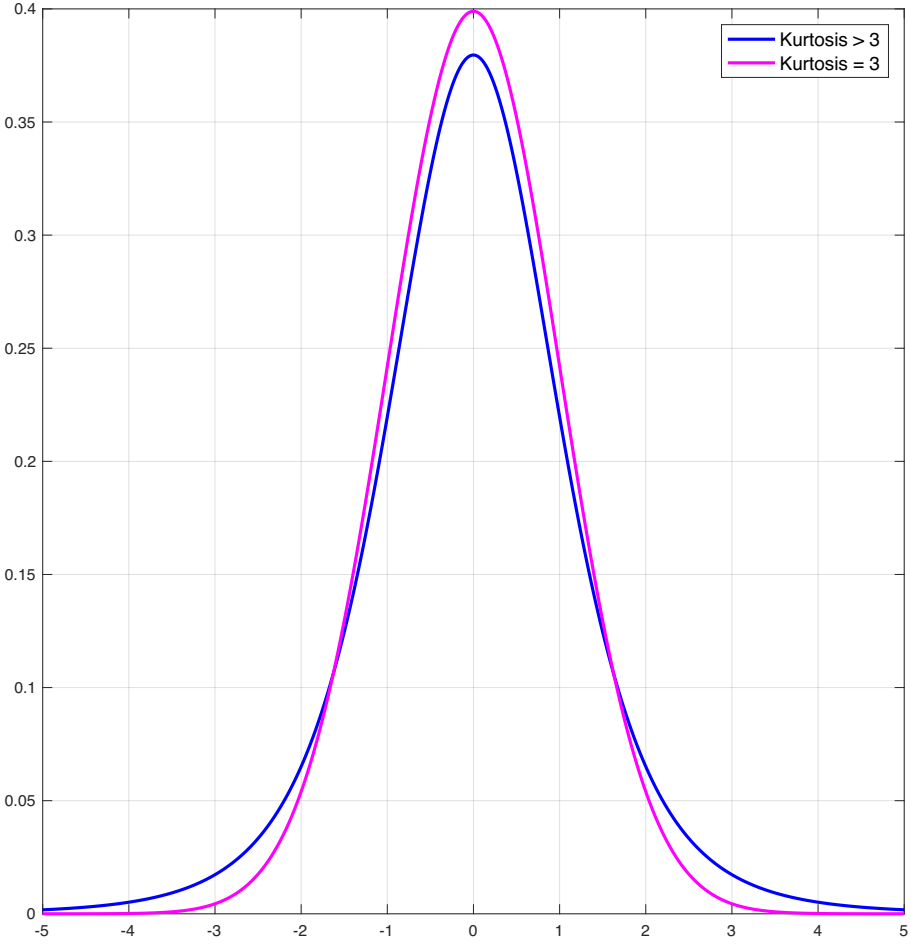
Properties of a random variable



Properties of a random variable

- **Kurtosis**
  - Measures the "thickness of the tails" of a distribution.
  - General formula: $\mathbb{E}\left[(X - \mathbb{E}[X])^4\right] / \sigma_X^4$
  - Unit free
  - Normal random variable: kurt $= 3$
  - The "benchmark" for kurtosis is often the kurtosis of a normal random variable.
    - Hence, we often talk about "excess" kurtosis, i.e., kurt $- 3$

Properties of a random variable



Properties of a random variable

	Bernoulli(p)	Uniform[0,1]	Uniform[a, b]	$N(0,1)$	$N(\mu_X, \sigma_X^2)$
PMF/PDF	$p^x(1-p)^{1-x}$	1	$\frac{1}{b-a}$	$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$	$\frac{1}{\sigma_X\sqrt{2\pi}} e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}}$
Mean	p	$\frac{1}{2}$	$\frac{1}{2}(a+b)$	0	μ_X
Median	$\begin{cases} 0 & \text{if } p < 1/2 \\ [0,1] & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$	$\frac{1}{2}$	$\frac{1}{2}(a+b)$	0	μ_X
Variance	$p(1-p)$	$\frac{1}{12}$	$\frac{1}{12}(b-a)^2$	1	σ_X^2
Stdev	$\sqrt{p(1-p)}$	$\sqrt{\frac{1}{12}}$	$\sqrt{\frac{1}{12}}(b-a)$	1	σ_X
Skew	$\frac{1-2p}{\sqrt{p(1-p)}}$	0	0	0	0
Kurt	$\frac{1-3p(1-p)}{p(1-p)}$	$\frac{9}{5}$	$\frac{9}{5}$	3	3
Ex. Kurt	$\frac{1-6p(1-p)}{p(1-p)}$	$-\frac{6}{5}$	$-\frac{6}{5}$	0	0

Other important probability distributions

- **Chi-squared distribution**
  - This is the distribution of the sum of squared independent standard normal random variables.
  - If $Z_i \sim N(0,1)$ then
$$\sum_{i=1}^{m} Z_i^2 \sim \chi_m^2$$
- **Student-t distribution**
  - If $Z \sim N(0,1)$ and $W \sim \chi_m^2$ (where $Z \perp\!\!\!\perp W$) then
$$\frac{Z}{\sqrt{W/m}} \sim t_m$$
  - When $m \geq 30$, the t-distribution is well approximated by a normal distribution.
    - A t_∞ distribution is equal to a standard normal distribution.
- **F distribution**
  - If $W \sim \chi_m^2$ and $V \sim \chi_n^2$ (where $W \perp\!\!\!\perp V$) then
$$\frac{W/m}{V/n} \sim F_{m,n}$$

Random sampling & sample average

- **Simple random sample**
  - For each draw, each object in the population is equally likely to be chosen.
  - Hence, the distribution is the same for all draws.
  - **Terminology**: each draw is said to be *identically distributed*.
  - Further, since each draw is independent of all other draws we expand our terminology and call the draws *independent and identically distributed*, which gets abbreviated to **iid**.
- **Sample average**
  - Suppose we have a population of size N and we draw a sample of size n (where $n < N$) from this population.
  - Now suppose we calculate the average over the sample

$$\bar{X} = \frac{1}{n} \underbrace{(X_1 + X_2 + \cdots + X_n)}_{\text{iid random draws}} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

  - $\bar{X}$ is called the **sample average**.

Sample average

- **Sample average**
  - Note that for any $n < N$, the sample average is itself a random variable.
  - Hence, the sample average has a distribution with a mean, a variance, etc.
  - What is $\mathbb{E}[\bar{X}]$?

$$\begin{aligned}
\mathbb{E}[\bar{X}] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu_X \\
&= \mu_X
\end{aligned}$$

Sample average

- **Sample average**
  - The equality $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i]$ follows by the linearity of the expectation operator, i.e.,

$$\mathbb{E}[a + bX + cY] = a + b\mathbb{E}[X] + c\mathbb{E}[Y]$$

    where X and Y are random variables and $a, b, c \in \mathbb{R}$.
  - The equality $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu_X$ follows from the fact that

$$\mathbb{E}[X_i] = \mu_X, \quad \forall i$$

    since each X_i is an iid draw from the population distribution and

$$\mu_X = \frac{1}{N}\sum_{i=1}^{N} X_i$$

    i.e., μ_X is the population mean.

Sample average

- **Sample average**
  - What is $\mathbf{V}ar(\bar{X})$?

$$\begin{aligned}
\mathbf{V}ar(\bar{X}) &= \mathbf{V}ar\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathbf{V}ar(X_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma_X^2 \\
&= \frac{\sigma_X^2}{n}
\end{aligned}$$

  - Note that

$$\frac{\sigma_X^2}{n} \to 0$$

    as $n \to N$ (for a finite population) or as $n \to \infty$ (for an infinite population).

Sample average

- **Sample average**
  - The equality $\mathbf{V}ar\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{V}ar(X_i)$ follows by independence. Recall that

    $$\mathbf{V}ar(a + bX + cY) = b^2\mathbf{V}ar(X) + c^2\mathbf{V}ar(Y) + 2bc\mathbf{C}ov(X, Y)$$

    where X and Y are random variables and $a, b, c \in \mathbb{R}$. If X and Y are independent (in which case $\mathbf{C}ov(X, Y) = 0$), then

    $$\mathbf{V}ar(a + bX + cY) = b^2\mathbf{V}ar(X) + c^2\mathbf{V}ar(Y)$$

  - The equality $\frac{1}{n^2}\sum_{i=1}^{n}\mathbf{V}ar(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_X^2$ follows from the fact that

    $$\mathbf{V}ar(X_i) = \sigma_X^2, \quad \forall i$$

    since each X_i is an iid draw from the population distribution and

    $$\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)^2$$

    i.e., σ_X^2 is the population variance.

Sample variance

- **Sample variance**
  - Given a data sample, we typically do no know the mean or variance of the population from which it was drawn.
  - A natural *estimator* for the population mean is the sample mean, $\bar{X}$.
  - A natural *estimator* for the population variance is the *sample variance*

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

  - Again note that for any $n < N$, the sample variance is itself a random variable.
  - Hence, the sample variance has a distribution with a mean and a variance.

Sample variance

- **Sample variance**
  - What is $\mathbb{E}[S_X^2]$?

$$
\begin{aligned}
\mathbb{E}[S_X^2] &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \\
&= \mathbb{E}\left(\frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]\right) \\
&= \frac{1}{n-1}(n\mathbb{E}[X_i^2] - n\mathbb{E}[\bar{X}^2]) \\
&= \frac{1}{n-1}\left(n(\sigma_X^2 + \mu_X^2) - n\left(\frac{\sigma_X^2}{n} + \mu_X^2\right)\right) \\
&= \sigma_X^2
\end{aligned}
$$

Sample variance

- **Sample variance**
  - The equality

$$\mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right]=\mathbb{E}\left(\frac{1}{n-1}\left[\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right]\right)$$

follows from the fact that

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\bar{X})^2 &= \sum_{i=1}^{n}(X_i^2-2X_i\bar{X}+\bar{X}^2)\\ &= \sum_{i=1}^{n}X_i^2-2\sum_{i=1}^{n}X_i\bar{X}+\sum_{i=1}^{n}\bar{X}^2\\ &= \sum_{i=1}^{n}X_i^2-2\bar{X}\sum_{i=1}^{n}X_i+n\bar{X}^2\\ &= \sum_{i=1}^{n}X_i^2-2n\bar{X}^2+n\bar{X}^2\\ &= \sum_{i=1}^{n}X_i^2-n\bar{X}^2\end{aligned}$$

Sample variance

- **Sample variance**
  - The equality

$$\frac{1}{n-1}(n\mathbb{E}[X_i^2]-n\mathbb{E}[\bar{X}^2]) = \frac{1}{n-1}\left(n(\sigma_X^2+\mu_X^2) - n\left(\frac{\sigma_X^2}{n}+\mu_X^2\right)\right)$$

follows from the facts that

$$\mathbf{V}ar(X_i) = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2$$

which implies

$$\mathbb{E}[X_i^2] = \sigma_X^2 + \mu_X^2$$

and

$$\mathbf{V}ar(\bar{X}) = \mathbb{E}[\bar{X}^2] - \mathbb{E}[\bar{X}]^2$$

which implies

$$\mathbb{E}[\bar{X}^2] = \frac{\sigma_X^2}{n} + \mu_X^2$$

Sample variance

- **Sample variance**
  - What is $\mathbf{V}ar(S_X^2)$?
  - This is a complicated expression in general, here we just note that (as with $\mathbf{V}ar(\bar{X})$)

$$\mathbf{V}ar(S_X^2) \to 0$$

    as $n \to N$ (for a finite population) or as $n \to \infty$ (for an infinite population).

Convergence of random variables

- Convergence of a sequence of random variables to some limit random variable.

- **Idea:** A sequence of random variables can sometimes be expected to settle down into a behaviour that is essentially unchanging when items far enough into the sequence are studied.

Convergence of random variables

- **Two basic notions:**
  - The sequence eventually takes a constant value.
  - Values in the sequence continue to change, but can be described by an unchanging probability distribution.
- **Our random variables**: statistics that are calculated based on random samples.
- **Our sequences**: defined by the size of the sample.

Convergence in probability

- **Idea:** The probability of an "unusual" outcome becomes smaller and smaller as the sequence progresses (i.e., as the sample size grows).

- **Formally:** A sequence of random variables $\{Y_n\}$ converges in probability to a constant value c if for any $\epsilon > 0$

$$\mathbb{P}(|Y_n - c| \leq \epsilon) \underset{n \to \infty}{\longrightarrow} 1,$$

or equivalently

$$\mathbb{P}(|Y_n - c| > \epsilon) \underset{n \to \infty}{\longrightarrow} 0$$

- **Notation:** $Y_n \xrightarrow{\mathbb{P}} c$ or $\operatorname{plim}_{n \to \infty} Y_n = c$

Convergence in probability

- **Consistency:** We call an estimator consistent if it converges in probability to the quantity being estimated.
- **Example:** $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a consistent estimator of μ_x if

$$\bar{X}_n \xrightarrow{\mathbb{P}} \mu_x,$$

where μ_x is the common mean of the X_i.

- **Example:** $S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a consistent estimator of σ_X^2 if

$$S_X^2 \xrightarrow{\mathbb{P}} \sigma_X^2,$$

where σ_X^2 is the common variance of the X_i.

Convergence in distribution

- A sequence of random variables $\{Y_n\}$ is said to converge in distribution to a random variable Y if [1]

$$\lim_{n\to\infty} F_n(c) = F(c),$$

for every $c \in \mathbb{R}$ at which F is continuous.

  - F_n is the CDF of Y_n (i.e., $F_n(c) = \mathbb{P}(Y_n \leq c)$), and
  - F is the CDF of Y (i.e., $F(c) = \mathbb{P}(Y \leq c)$)

- **Notation:** $Y_n \xrightarrow{d} Y$

- If Y is a standard normal random variable (i.e., $Y \sim N(0,1)$), we also use the notation $Y_n \xrightarrow{d} N(0,1)$.

[1] Our convention in this class will be to use $f(x)$ to refer to the PMF/PDF of a distribution and to use $F(c) = \int_{-\infty}^{c} f(x)dx$ to refer to the CDF of the distribution.

Markov inequality

- If X is a random variable with $\mathbb{P}(X > 0) = 1$, then for any $c > 0$

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}[X]}{c}$$

- **Proof:**
  - We can write

$$\mathbb{E}[X] = \int_0^\infty x f(x) dx = \int_0^c x f(x) dx + \int_c^\infty x f(x) dx$$

  - hence

$$\mathbb{E}[X] \geq \int_c^\infty x f(x) dx \geq c \int_c^\infty f(x) dx = c\mathbb{P}(X \geq c)$$