Central Dogma

Probability Notation
Probability identities \( \leq \) start here.

Statistics

- mean \( E[x] \)
- Variance
- Covariance

Mechanism

Least time.
- today
Probabilistic Truths

\[ P(A \text{ or } B) = P(A) + P(B) \text{ if } A \text{ and } B \text{ are mutually exclusive}. \]

For any events \( A \) and \( B \):
\[ P(A) = P(A, B) + P(A, \neg B) \]
\( b/c \) "\( A \) and \( B \)" and "\( A \) and \( \neg B \)" are mutually exclusive.

More generally, for any set of events \( B_1, B_2, \ldots, B_n \) such that exactly one of \( B_i \) must be true:
\[ P(A) = P(A, B_1) + P(A, B_2) + \ldots + P(A, B_n) \]
This is the law of total probability.

Calculating \( P(A) \) by summing its probs over all \( B_i \)'s is called marginalizing and \( P(A) \) is called the marginal distribution of \( A \).

If we know prob of \( B \) and prob of \( A \) given \( B \), we can determine prob of \( A \) and \( B \):
\[ P(A, B) = P(A|B) \cdot P(B) \implies P(A|B) = \frac{P(A, B)}{P(B)} \]
\( \Rightarrow \) this also gives us a numerical representation of independence: \( P(A, B) = P(A) \cdot P(B) \)

\( \Rightarrow \) And combining this w/ symmetry relation \( P(A, B) = P(B, A) \)
gives us something very important.

\[ P(A, B) = P(A|B) \cdot P(B) = P(B, A) = P(B|A) \cdot P(A) \]
\[ P(A|B)P(B) = P(B|A)P(A) \]
\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]

Bayes' Rule (Bayes' Thm)

\[ \text{[We will use Bayes' Thm. in the future, quite a bit]} \]

Lastly, \( P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots + P(A|B_n)P(B_n) \)
combining law of total prob w/ conditional probs.
Statistics. A statistic is a (numerical) measure of a "feature" of a probability distribution.

- **Expected Value** aka "mean" (expectation), can be used for variables w/ numerical values.

The expected value $E[X]$ of a variable $X$ is:

$$E[X] = \sum_x x P(X=x)$$

More generally, for a function of $X$:

$$E[g(X)] = \sum_x g(x) P(X)$$

(Example: $g(x) = x^2$)

**Conditional Expectation:**

$$E[Y|X=x] = \sum_y y P(Y=y|X=x)$$

$E[X]$ is useful for making "guesses" of $X$'s value, b/c $f = E[X]$ is the function which minimizes the expected square error $E[(f-X)^2]$.

And $E[Y|X=x]$ is a best guess of $Y$ given we have observed $X=x$, b/c $g = E[Y|X=x]$ minimizes $E[(g-Y)^2|X=x]$

(This assumes implicitly that $P(X)$ or $P(Y|X=x)$ is approximately symmetric. If these distributions are skewed, it's better to use other statistics, such as the median, which minimizes the expected absolute error.)

Guessing and error measures are closely related to loss functions in ML/SL and optimization problems more generally.
Variance: \( \text{Var}(X) = E[(X - \overline{X})^2] \), \( \overline{X} = E[X] \)

\[
= E[(X - E[X])^2] \\
= E[X^2 - 2E[X]X + E[X]^2] \\
= E[X^2] - 2E[X]^2 + E[X]^2 \\
= E[X^2] - E[X]^2 \\
\]

Covariance: \( \text{cov}(X,Y) = E[(X-\overline{X})(Y-\overline{Y})] \)

\[
\text{cov}(X,Y) = 0 \text{ if } X \text{ and } Y \text{ are uncorrelated} \\
\]

Standard deviation \( \sigma_X = \sqrt{\text{Var}(X)} \), sometimes \( \text{Var}(X) = \sigma_X^2 \)

S.D. is nice since it has the same units as the original variable
Properties of mean and variance

1. Expectation is a linear function
   • \( E[c_i X + c_a Y] = c_i E[X] + c_a E[Y] \), \( c_i, c_a \) const.
   • \( E\left[ \sum_{i=1}^{n} c_i X_i \right] = \sum_{i=1}^{n} c_i E[X_i] \) in general

2. Variance obeys:
   • \( \text{Var}(X) \geq 0 \) w/ equality iff random variable is a constant
   • \( \text{Var}(X + c) = \text{Var}(X) \)
   • \( \text{Var}(cX) = c^2 \text{Var}(X) \)
   • \( \text{Var}\left( \sum_{i=1}^{n} c_i X_i \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \text{Cov}(X_i, X_j) \)
     \[ = \sum_{i=1}^{n} c_i^2 \text{Var}(X_i) + \sum_{i \neq j} c_i c_j \text{Cov}(X_i, X_j) \]
   • If R.V.s \( X_i \) are \underline{uncorrelated}:
     \( \text{Var}\left( \sum_{i=1}^{n} c_i X_i \right) = \sum_{i=1}^{n} c_i^2 \text{Var}(X_i) \)
     \( \text{w/C} \) \( \text{Cov}(X_i, X_j) = 0 \) if \( i \neq j \)