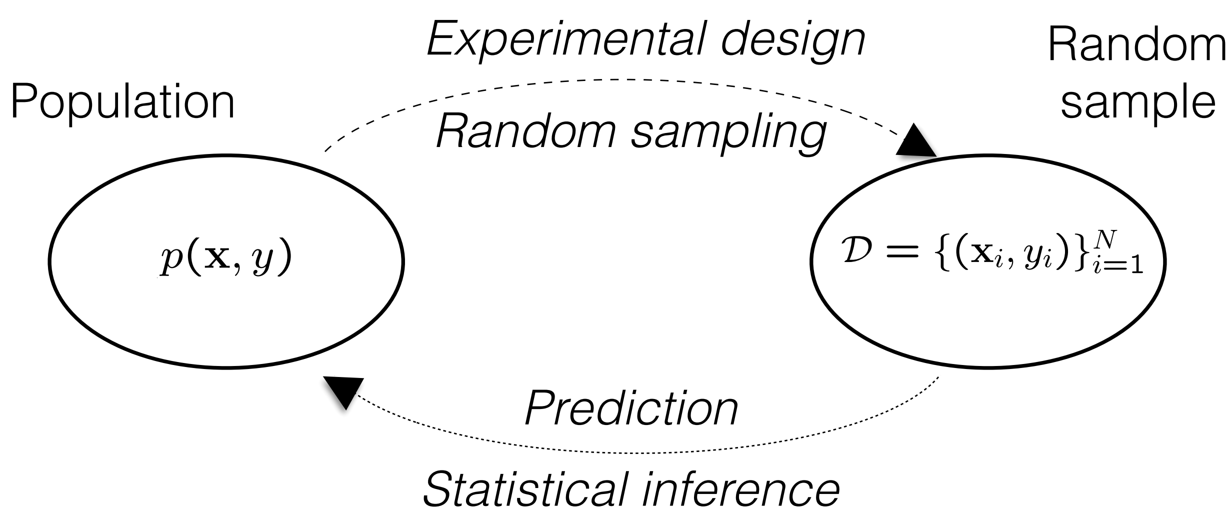
Probability distributions

Kevin Murphy Ch. 2

CS 6140
Machine Learning
Professor Olga Vitek

January 12, 2017

Random sampling



Probabilistic statements

$p(y \leq a|\mathbf{x})$ $\quad$ $p(\bar{y} \leq a|\mathbf{x})$

Model-based summaries

$\hat{y} = \hat{f}(\mathbf{x})$ $\quad$ $p(y|\mathbf{x}, \mathcal{D})$

Probability: frequentist interpretation

- A thought experiment
- Probability of an outcome
  - Defined in the context of chance operation
  - Quantifies the chance of occurrence of the event
- Probability calculations
  - The proportion of times that the outcome will occur in an infinite sequence of observations

$$P\{E\} = \frac{\#\text{ favorable outcomes}}{\#\text{ possible outcomes}}$$

- Describes what the outcomes would be
  - If the population parameters were known
  - If we could measure the population

Probability: Bayesian interpretation

- Uncertainty in in event
  - Including events that do not have long-term frequencies (e.g., ice cap melts)
  - Relates to information instead of repeated trials
- Can express our subjective uncertainty
- Describes what the outcomes would be
  - If the population parameters were known
  - If we could measure the population

Properties

- Axioms of probabilities
  - Probability of empty set $P\{\text{No event}\} = 0$
  - Probability of any event $P\{\text{Any event}\} = 1$
  - Probability of union of two events
    $P\{E_1 \cup E_2\} = P\{E_1\} + P\{E_2\} - P\{E_1 \cap E_2\}$

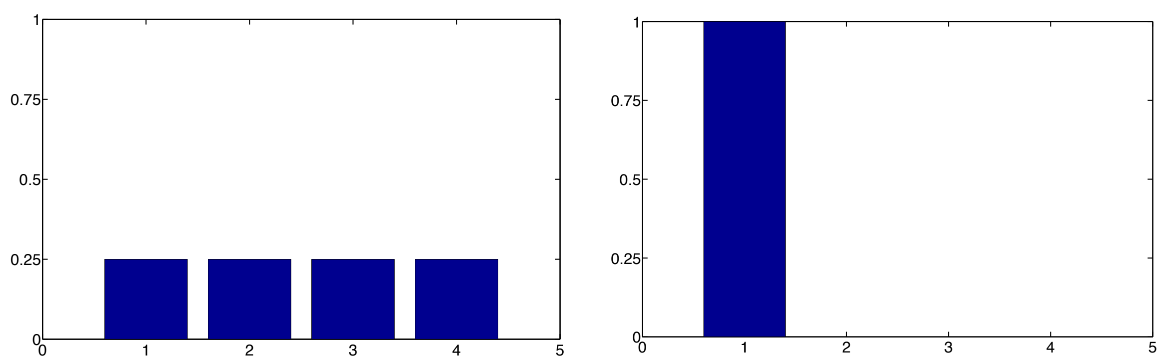
- Consequences
  - $0 \leq P\{E\} \leq 1$
  - $P\{\bar{E}\} = 1 - P\{E\}$
  - Monotonicity If $E_1 \in E_2$, then $P\{E_1\} \leq P\{E_2\}$
  - If $E_1 \cup E_2 \leq P\{E_1\} + P\{E_2\}$
  - If $E_1 \cap E_2 = \emptyset$, then $P\{E_1 \cup E_2\} = P\{E_1\} + P\{E_2\}$

Categorical random variables Probability distribution

- A table
  - Or a formula describing the table

- Frequencies of events in the population
  - Coin toss $P\{\text{Tail}\} = \dfrac{\#\ \text{Tail}}{\#\ \text{Tail} + \text{Head}}$
  - Lottery win $P\{\text{Win}\} = \dfrac{\#\ \text{Winning tickets}}{\#\ \text{All tickets}}$

- Bayesian: confidence in the event



K. Murphy, Fig 2.1

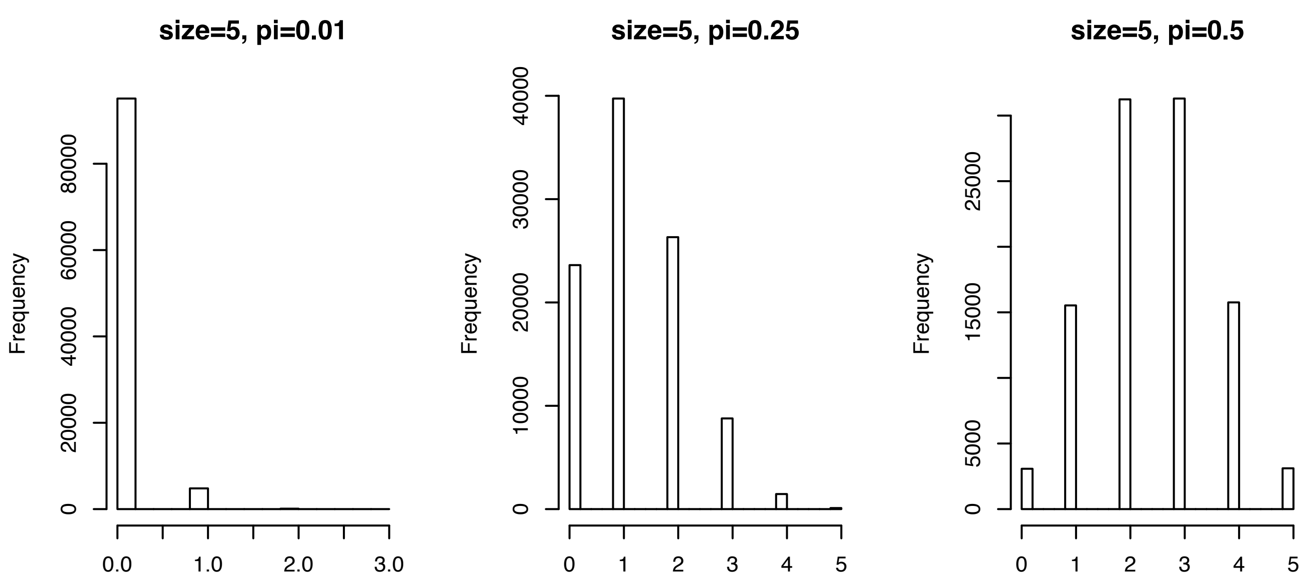
Examples

- Bernoulli(π) - parameter π

$$P\{Y = y\} = \begin{cases} \pi, & \text{if } y = 1 \\ 1 - \pi, & \text{if } y = 0 \end{cases}$$

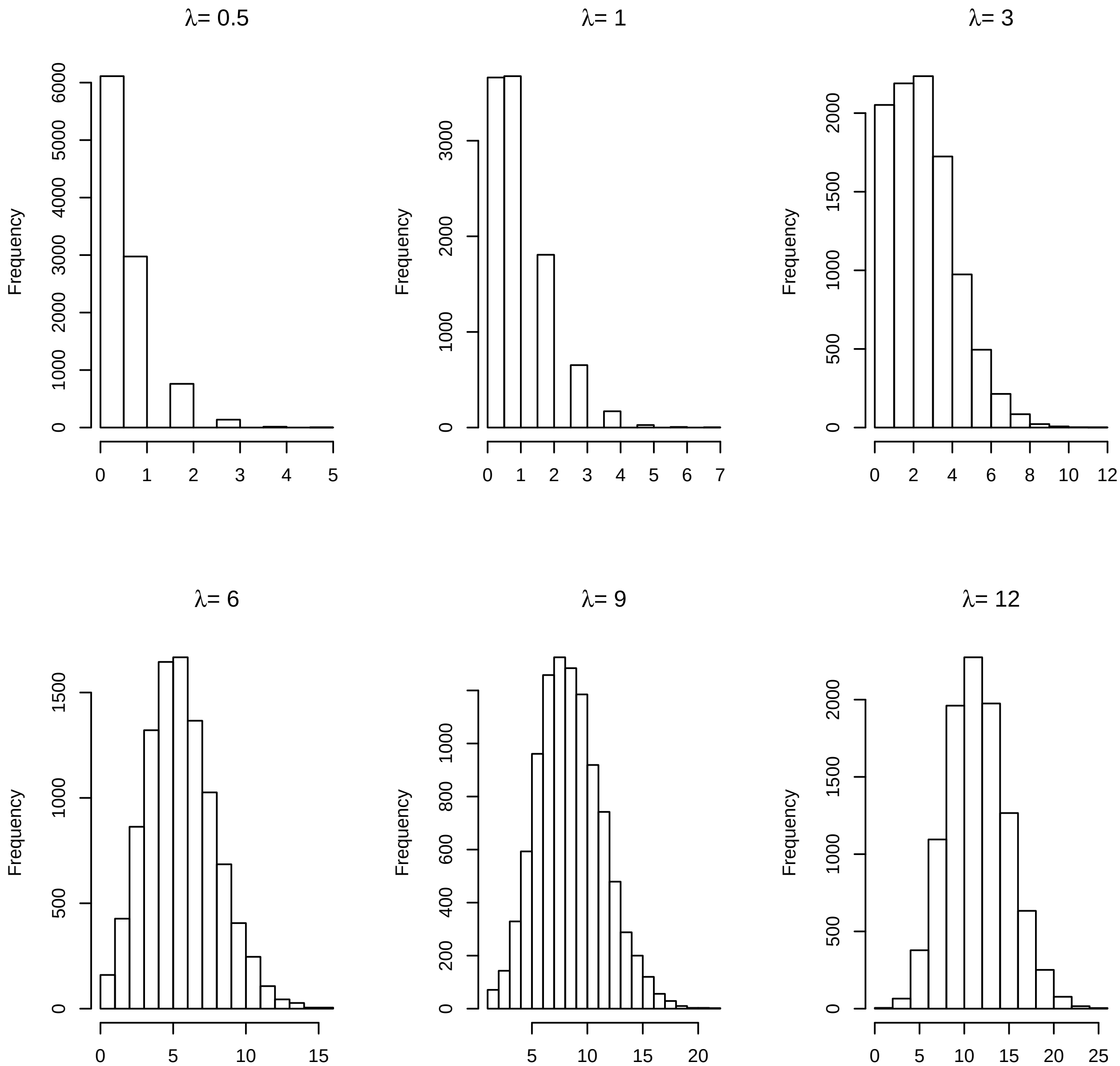
- Binomial(n, π) - parameters (n, π)

$$P\{Y = y\} = \binom{n}{y}\pi^y(1-\pi)^{n-y},\ y = 0, \ldots n$$



Poisson(λ)

$$P\{Y = y\} = \frac{e^{-\lambda}\lambda^y}{y!}, \; y = 0, \ldots$$



Probability distribution: two categorical random variables

- Joint outcome of two events
  - Tossing two coins

	Head	Tail
Head	$\frac{1}{4}$	$\frac{1}{4}$
Tail	$\frac{1}{4}$	$\frac{1}{4}$

  - $P\{\text{Coin}_1 = \text{Head and Coin}_2 = \text{Tail}\} = \frac{1}{4}$

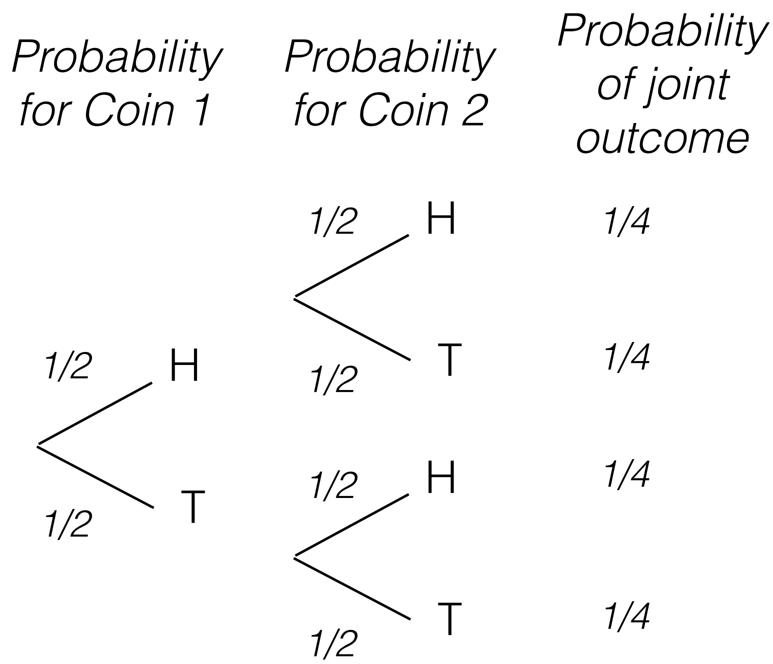
- Can think of this as a single event with a bivariate pattern

Head,Head	Head,Tail	Tail,Head	Tail,Tail
$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

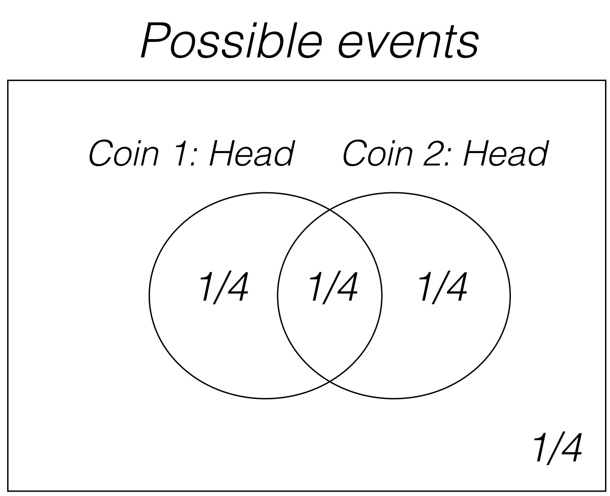
- More useful to think of this as two distinct events, and study their joint properties

Joint outcome of two events

- Probability tree



- Venn diagram



Outcomes of two events

- Outcome of two events
  - $P\{E_1, E_2\} = P\{$(Head, Head)$\} = 1/4$
  - $P\{E_1, E_2\} = P\{$(Head, Tail)$\} = 1/4$
  - $P\{E_1, E_2\} = P\{$(Tail, Head)$\} = 1/4$

- Outcome of any of the two events
  - $P\{E_1 \text{ OR } E_2\} = P\{E_1 \cup E_2\} =$
    $P\{E_1\} + P\{E_2\} - P\{E_1 \cap E_2\}$
  - $P\{$Coin 1 Head OR Coin 2 Head$\} =$
    $2/4 + 2/4 - 1/4 = 3/4$
  - $P\{$At least 1 Head$\} =$
    $2/4 + 2/4 - 1/4 = 3/4$
  - $P\{$(Head, Tail), any order$\} =$
    $P\{$Head, Tail$\} + P\{$Tail, Head$\} - 0 = 1/4 + 1/4$

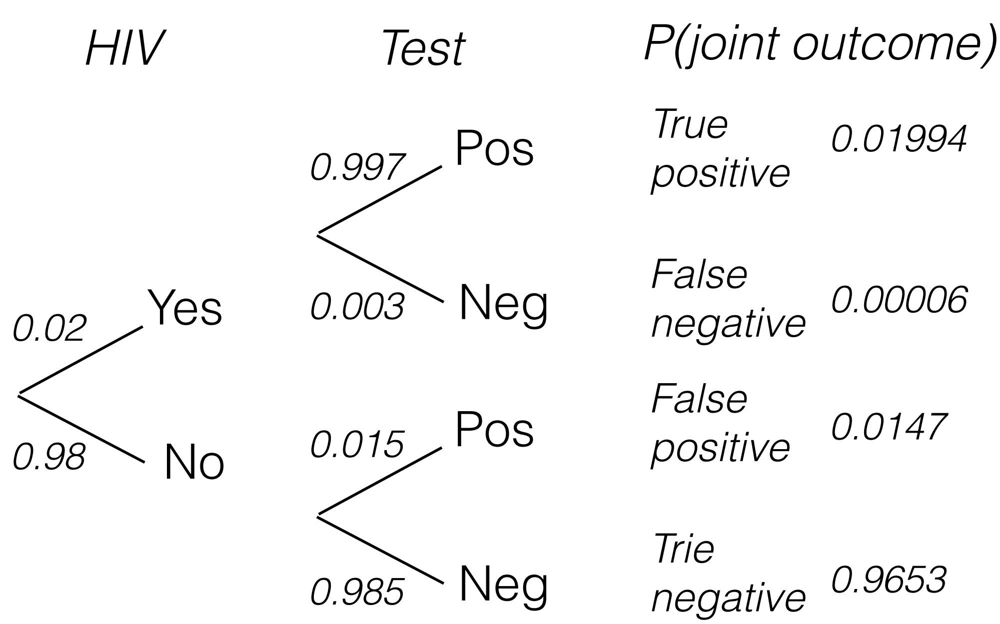
Conditional probability and Bayes rule

- Conditional probability
  - $P\{E_2|E_1\} = \frac{P\{E_1 \ \cap \ E_2\}}{P\{E_1\}}$
  - Example:
    $P\{\text{Coin 2 Head} \mid \text{Coin 1 Head}\} =$
    $\frac{P\{\text{Coin 2 Head} \cap \text{Coin 1 Head}\}}{P\{\text{Coin 1 Head}\}} = \frac{1/4}{1/2} = 1/2$
- Consequence
  - $P\{E_1 \ \cap \ E_2\} = P\{E_1\} \cdot P\{E_2|E_1\}$
  - Used to draw probability trees
- Bayes rule

$$\begin{aligned} P\{E_2|E_1\} &= \frac{P\{E_1 \ \cap \ E_2\}}{P\{E_1\}} = \frac{P\{E_2\} \cdot P\{E_1|E_2\}}{P\{E_1\}} \\ &= \frac{P\{E_2\} \cdot P\{E_1|E_2\}}{\sum_{e_2} P\{e_2\} \cdot P\{E_1|e_2\}} \end{aligned}$$

Example

- HIV testing
  - The prevalence of HIV in a population is 2%
  - Test for HIV has sensitivity 99.7%
  - The test is negative in 98.5% of healthy people
- What is the probability that the person with a positive test has HIV?



Generalization

- Goal: classify $y \in \{1, \ldots, C\}$
  - Bayes rule:

$$p(y = c|\mathbf{x}) = \frac{p(y = c) \cdot p(\mathbf{x}|y = c)}{\sum_{c'} p(y = c') \cdot p(\mathbf{x}|y = c')}$$

- Generative classifiers
  - Specify prior probability of $p(y = c)$
  - Assume class-conditional distribution $p(\mathbf{x}|y = c)$
  - Use Bayes rule to derive the posterior $p(y = c|\mathbf{x})$
  - **Example:** Linear discriminant analysis

- Discriminative classifiers
  - Estimate the posterior the posterior $p(y = c|\mathbf{x})$
  - Do not assume the distribution on $\mathbf{x}$
  - **Example:** Logistic regression

Independence

- Independence
  - Two events are independent if $P\{E_2|E_1\} = P\{E_2\}$
  - Example:

    $P\{\text{Coin 2 Head} \mid \text{Coin 1 Head}\} = P\{\text{Coin 2 Head}\}$
    Therefore two coins are independent
  - Consequence: for independent events

    $P\{E_1 \cap E_2\} = P\{E_1\} \cdot P\{E_2|E_1\} = P\{E_1\} \cdot P\{E_2\}$

- Example: hair versus eye color
  - Assume that we have measurements on the entire population of individuals

Eye color	Hair color: Brown	Black	Red	Total
Brown	400	300	20	720
Blue	800	200	50	1050
Total	1200	500	70	1770

  - Are hair and eye colors independent?

Continuous random variables Probability density function

- Probability density function (pdf)
  - An idealized histogram of Y
  - Continuum paradox: for any a, $P\{Y = a\} = 0$

- Cumulative distribution function (cdf)
  - $F(a) = P\{y \leq a\}$
  - Probability that Y is between a and b
P$\{\ a \leq Y \leq b\ \} = F(b) - F(a)$
  - $F(a) = P\{y \leq a\} = P\{y < a\}$

- Numbers such as a and b are called *quantiles* of the probability density
  - E.g., a is the 25th quantile, if P$\{Y \leq a\} = 0.25$

Examples

- The Uniform distribution $\mathcal{U}(a, b)$

  $$f(Y) = \begin{cases} \frac{1}{b-a}, & \text{if } Y \in [a, b] \\ 0, & \text{otherwise} \end{cases}$$

- The general Normal distribution $\mathcal{N}(\mu, \sigma)$

  $$f(Y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{Y-\mu}{\sigma}\right)^2}$$

  – μ and σ are parameters

- The Standard Normal distribution $\mathcal{N}(0, 1)$

  $$f(Y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Y^2}$$

  – A α quantile of $Z \sim \mathcal{N}(0, 1)$ is z_α, such that $P\{Z \leq z_\alpha\} = \alpha$

Areas under $\mathcal{N}(0, 1)$

Standard Normal Probabilities

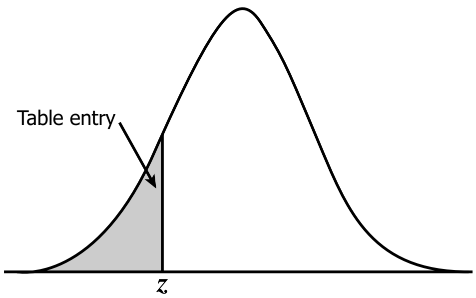


Table entry for z is the area under the standard normal curve to the left of z.

z	**.00**	**.01**	**.02**	**.03**	**.04**	**.05**	**.06**	**.07**	**.08**	**.09**
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559

```
> qnorm(0.025)
[1] -1.959964
> pnorm(1.96)
[1] 0.9750021
```

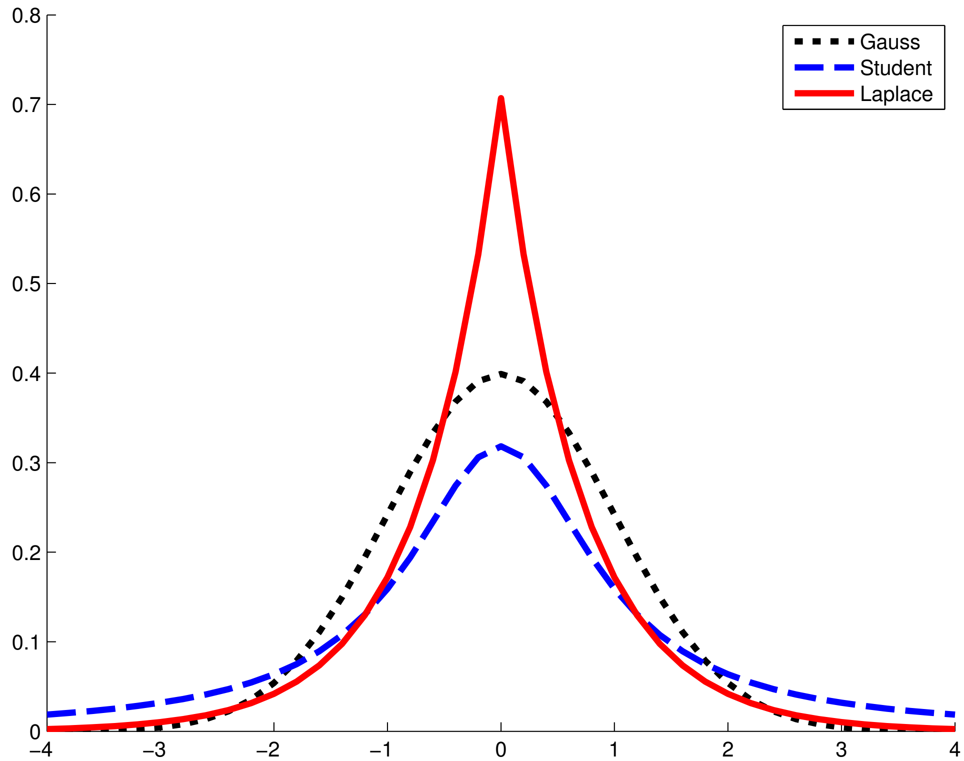
Examples

- The Student t distribution

  $$f(Y) \propto \left[+ \frac{1}{\nu} \left(\frac{Y - \mu}{\sigma} \right)^2 \right]^{-\frac{\nu+1}{2}}$$

- The Laplace distribution

  $$f(Y) = \frac{1}{2b} \exp \left(- \frac{|Y - \mu|}{b} \right)$$



K. Murphy, Fig 2.7

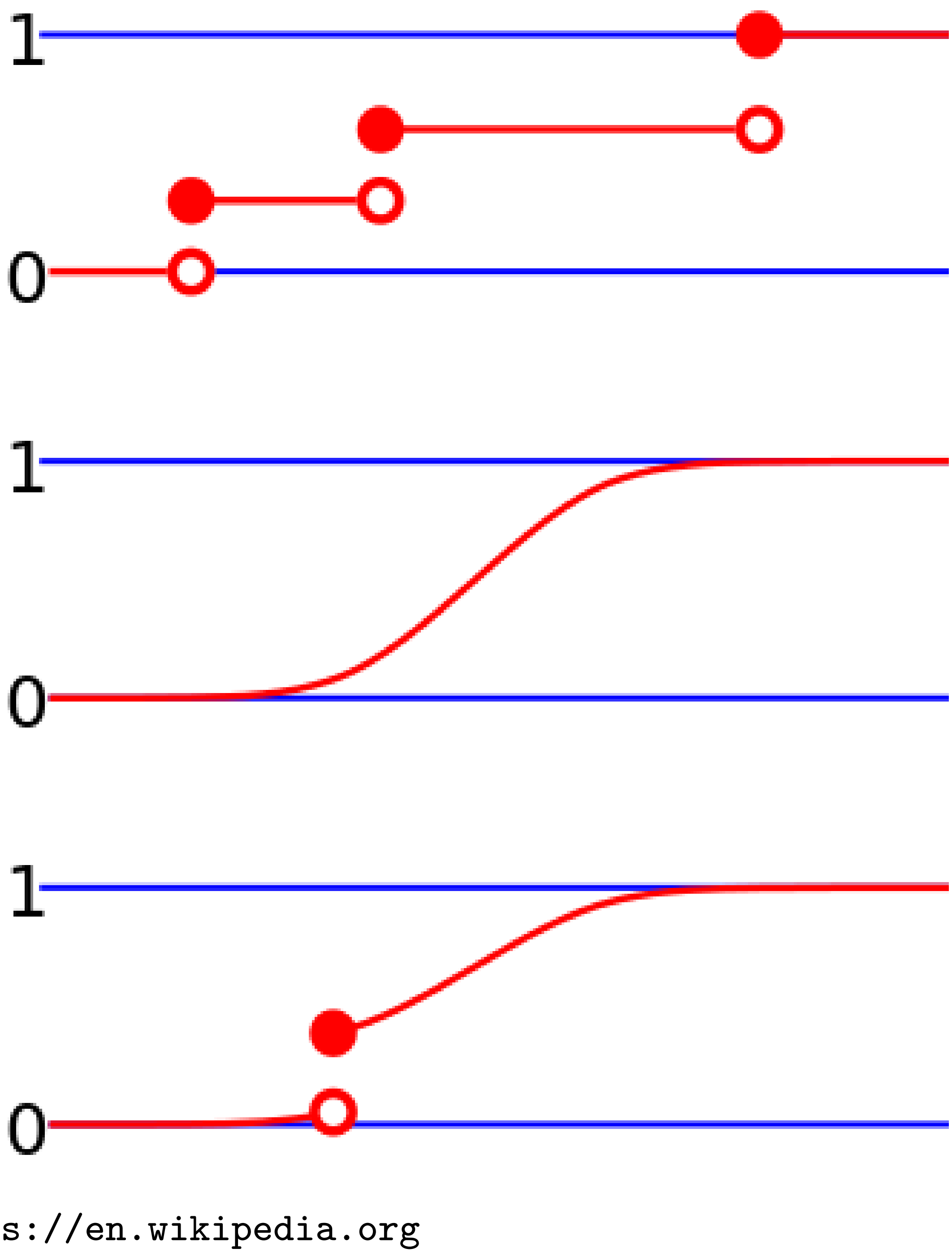
More on cumulative distribution function (CDF)

- $F_Y(y) = P\{Y \leq y\}$
  - Ordered categor. variables: $F_Y(y) = P\{Y \leq y\}$
  - Continuous variables: $F_Y(y) = \int\limits_{-\inf}^{x} f_Y(y)dx$

- Properties
  - $F_Y(y)$ is non-decreasing and right-continuous
  - $\lim_{y \to -\inf} F_Y(y) = 0$ and $\lim_{y \to \inf} F_Y(y) = 1$
  - y is the $F_Y(y)$th quantile of Y
  - $P\{a \leq X \leq b\} = F_Y(b) - F_Y(a)$
  - $\bar{F}_Y(y) = P\{Y > y\} = 1 - F_Y(y)$

- Probability density function: definition
  - $f_Y(y) = \frac{\partial F_Y(y)}{\partial y}$

Example



https://en.wikipedia.org

Expected value

- Expected value = population mean

- Denoted $E\{Y\} = \mu_Y$

- Ordered categorical variables
  - $E\{Y\} = \mu_Y = \sum_i y_i \, P\{Y = y_i\}$
  - The sum is over all possible values

- Continuous variables
  - $E\{Y\} = \mu_Y = \int_{-\text{inf}}^{\text{inf}} y f(y) dy$

- If X & Y have expected values μ_X and μ_Y:
  - $\mu_{aX+b} = a\mu + b$
  - $\mu_{X+Y} = \mu_X + \mu_Y$ and $\mu_{X-Y} = \mu_X - \mu_Y$
  - $\mu_{X/Y} \neq \mu_X / \mu_Y$

Variance

- Denoted $Var\{Y\} = \sigma^2\{Y\}$
- $\sigma_Y^2 = E\{(Y - E\{Y\})^2\}$
- Ordered categorical variables
  - $\sigma_Y^2 = \sum_i (y_i - \mu_y)^2 P\{Y = y_i\}$
  - The sum is over all possible values
- Continuous variables
  - $\sigma_Y^2 = \mu_Y = \int_{-\text{inf}}^{\text{inf}} (y - E\{Y\})^2 f(y)dy$
- If X & Y have expected values μ_X and μ_Y:
  - $\sigma_Y^2 aX + b = a^2\mu$
  - If X and Y are *independent* random variables $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$ and $\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2$
  - $\mu_{X/Y} \neq \mu_X/\mu_Y$

Examples

- $Y \sim$ Bernolli(π)
  - $E\{Y\} = \pi$, $Var\{Y\} = \pi(1 - \pi)$

- $Y \sim$ Binomial(n, π)
  - $E\{Y\} = n\pi$, $Var\{Y\} = n\pi(1 - \pi)$

- $Y \sim$ Poisson(λ)
  - $E\{Y\} = \lambda$, $Var\{Y\} = \lambda$

- $Y \sim \mathcal{U}(a, b)$
  - $E\{Y\} = \frac{a+b}{2}$, $Var\{Y\} = \frac{1}{12}(b - a)^2$

- $Y \sim \mathcal{N}(\mu, \sigma)$
  - $E\{Y\} = \mu$, $Var\{Y\} = \sigma^2$

- $Y \sim Student(\mu, \sigma, \nu)$
  - $E\{Y\} = \mu$, $Var\{Y\} = \frac{\nu\sigma^2}{(\nu-2)}$

Examples

- In a population of fish, the distribution of the number of tail vertebrate Y is

No. vertebrae	20	21	22	23	Total
% of fish	3	51	40	6	100

  - What is the population mean and variance of the distribution?

- For a random variable $Y \sim \mathcal{N}(\mu, \sigma)$:

$$\begin{aligned} P\{Y \leq y\} &= P\left\{\frac{Y-\mu}{\sigma} \leq \frac{y-\mu}{\sigma}\right\} \\ &= P\left\{Z \leq \frac{y-\mu}{\sigma}\right\} \end{aligned}$$

  - Use a table lookup (or R) for the standard Normal distribution

Joint probability distribution of two continuous variables

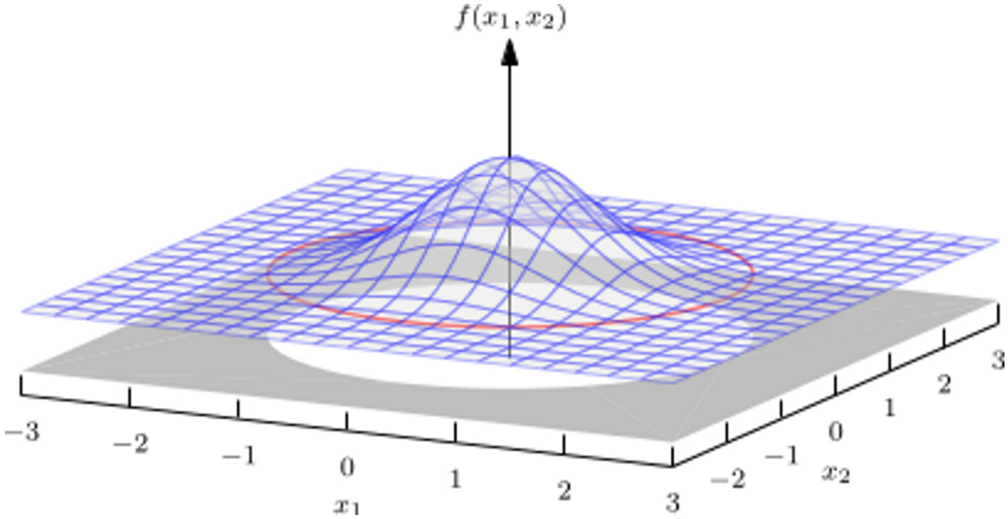
- Bivariate Normal distribution of (Y_1, Y_2)
- Requires five parameters
  - μ_1 and σ_1 are the mean and std dev of Y_1
  - μ_2 and σ_2 are the mean and std dev of Y_2
  - ρ_{12} is the coefficient of correlation

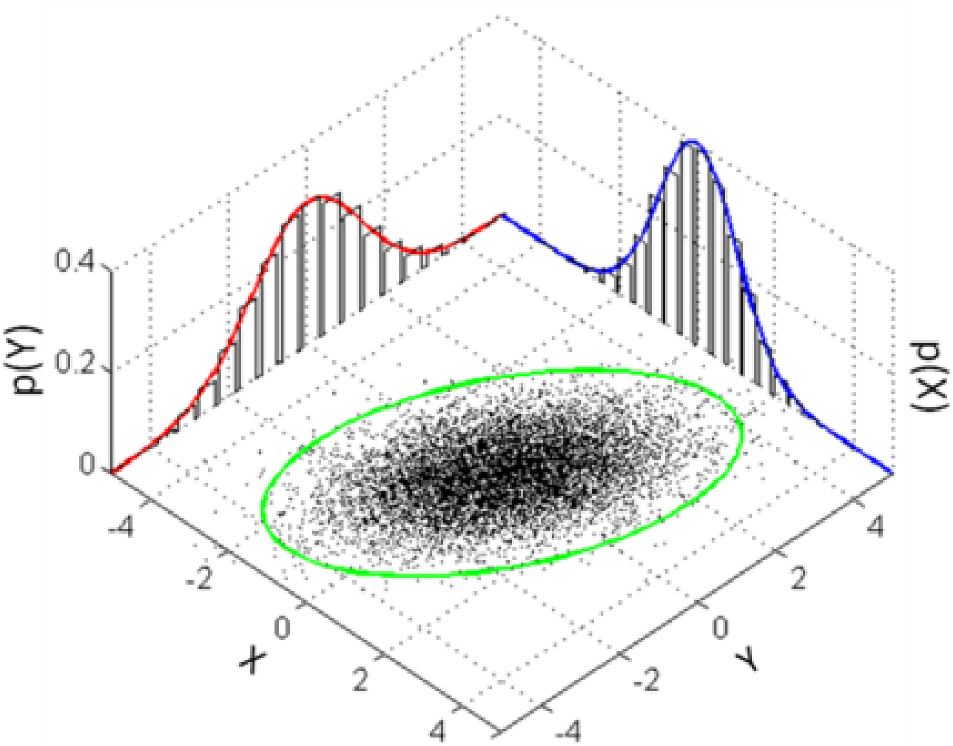
$$\begin{aligned}\rho_{12} &= \frac{Cov(Y_1, Y_2)}{\sigma_1\,\sigma_2} = \frac{E\{(Y_1-\mu_1)(Y_2-\mu_2)\}}{\sigma_1\,\sigma_2} \\ &= \frac{E\{Y_1\,Y_2\} - E\{Y_1\}E\{Y_2\}}{\sigma_1\,\sigma_2}\end{aligned}$$

- Bivariate normal density

$$\begin{aligned}f(Y_1, Y_2) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{12}^2}}\,\text{exp}\left\{-\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{Y_1-\mu_1}{\sigma_1}\right)^2\right.\right. \\ &\quad \left.\left. - -2\rho_{12}\left(\frac{Y_1-\mu_1}{\sigma_1}\right)\left(\frac{Y_2-\mu_2}{\sigma_2}\right)+\left(\frac{Y_2-\mu_2}{\sigma_2}\right)^2\right]\right\}\end{aligned}$$

Examples





https://en.wikipedia.org

Covariance and correlation

- Covariance: extent of linear relationship
  Correlation: normalized to [-1, 1]

- For two random variables X_1 and X_2

  $Cov(X_1, X_2) = E\{X_1 - E\{X_1\}\} \cdot E\{X_2 - E\{X_2\}\}$

  $Cor(X_1, X_2) = \frac{Cov(X_1,X_2)}{\sqrt{Var(X_1) \cdot Var(X_2)}}$

- For D-dimensional vectors

$$
\begin{aligned}
Cov(\mathbf{x}) &= E\left[(\mathbf{x} - E\{\mathbf{x}\})(\mathbf{x} - E\{\mathbf{x}\})'\right] = \\
&= \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_D) \\ Cov(X_2, X_1) & Var(X_2) & \cdots & Cov(X_2, X_D) \\ & & \cdots & \\ Cov(X_d, X_1) & Cov(X_d, X_2) & \cdots & Var(X_D) \end{pmatrix}
\end{aligned}
$$

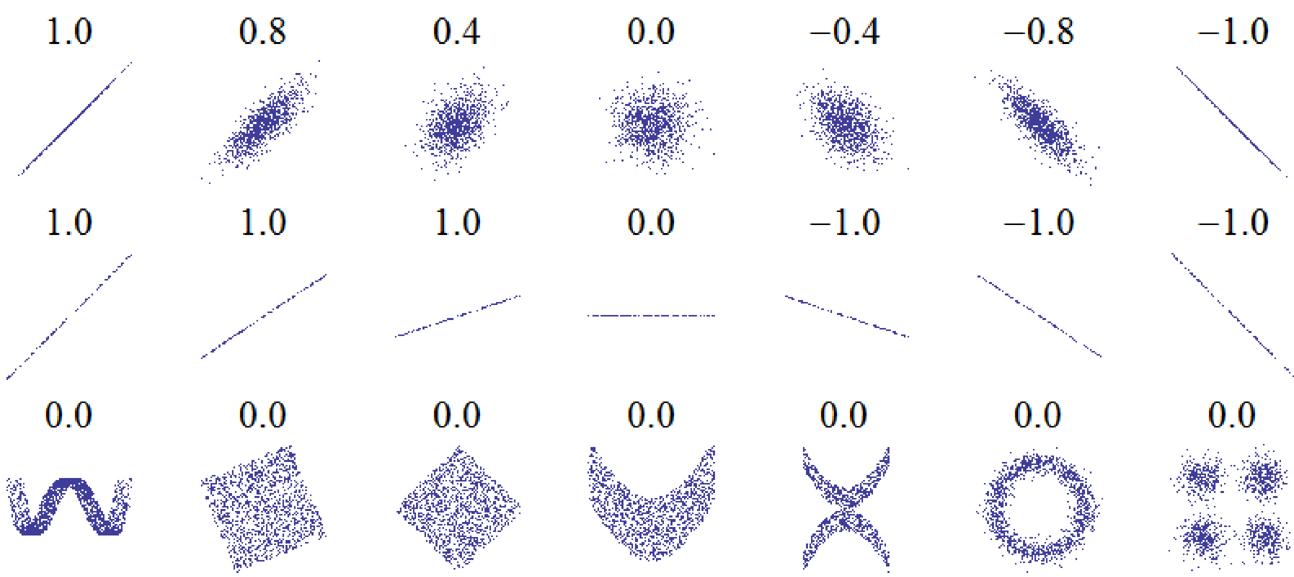
- Joint D-dimensional distribution $\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma})$

  $f(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu)'\mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)\right]$

Examples

- Uncorrelated r.v. $\neq$ independent r.v.!
  - Independent random variables are uncorrelated
  - Uncorrelated random variables may be dependent
  - Exception: Normal distribution is uncorrelated iff independent

- Examples of correlation



K. Murphy, Fig 2.12

Conditional Distribution

- Consider the distribution of Y_1 given Y_2

- Can show that $Y_1|Y_2$ is Normally distributed

- The mean can be expressed

$$E\{Y_1|Y_2\} = \left(\mu_1 - \mu_2\rho_{12}\frac{\sigma_1}{\sigma_2}\right) + \rho_{12}\frac{\sigma_1}{\sigma_2}Y_2 \;\; = \;\; \alpha_{1|2} + \beta_{12}Y_2$$

$\rightarrow$ Therefore $\beta_{1|2} = \rho_{12}\frac{\sigma_1}{\sigma_2}$

- With constant variance

$$Var\{Y_1|Y_2\} = \sigma_1^2\left(1 - \rho_{12}^2\right)$$

Information theory

- Entropy of a r.v. y with distribution p
  - Measure of its uncertainty
  - For a discrete variable with C states

    $$H(y) = -\sum_{c=1}^{C} p(y=c) \cdot log_2 p(y=c)$$

- Kullback-Leibler divergence (= rel. entropy)
  - Dissimilarity of two prob. distributions p and q
  - For discrete probability distributions with C states

$$\begin{aligned} KL(p,q) &= \sum_{c=1}^{C} p_c \cdot log_2 \frac{p_c}{q_c} = \sum_{c=1}^{C} p_c \, log_2 \, p_c - \sum_{c=1}^{C} p_c \, log_2 \, q_c \\ &= -H(p) + H(p,q) = \text{-entropy + cross-entropy} \end{aligned}$$

- Mutual information
  - General approach: associations between two r.v.
  - KL dissimilarity between joint distribution and product of marginal distributions
  - MI=0 iff the variables are independent

    $$I(x,y) = \sum_x \sum_y p(x,y) \, log_2 \, \frac{p(x,y)}{p(x)p(y)}$$

Sampling distribution

- Sampling variability
  - Variability of random samples from the population

- Sampling distribution
  - Variability of summaries of random samples from the population
  - Sampling distribution of the sample mean, sample variance, etc

- E.g.: sampling distribution of $\bar{Y}$

  1 Collect values $y_1, y_2, \ldots, y_n$ from the population

  2 Calculate the mean $\bar{y}$

  3 Repeat [1-2] a very large number of times, say 1,000,000

  4 The histogram of 1,000,000 values of $\bar{y}$ approximates the sampling distribution of $\bar{Y}$

The Central Limit Theorem

- The Central Limit Theorem
  - If $y_1, y_2, \ldots, y_n$ follow an arbitrary probability distribution with expected value μ and standard deviation σ, and n is large
  - then $\bar{Y} \overset{approximately}{\sim} \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

- Special case: $y_i \sim$ Normal distribution

  If $y_1, y_2, \ldots, y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma) \ \rightarrow \ \bar{Y} \overset{exactly}{\sim} \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

- Special case: $y_i \sim$ Bernoulli distribution

  If $y_1, y_2, \ldots, y_n \overset{iid}{\sim}$ Bernoulli $(\pi) \rightarrow \sum_{i=1}^{n} y_i \sim$ Binomial (n, π)

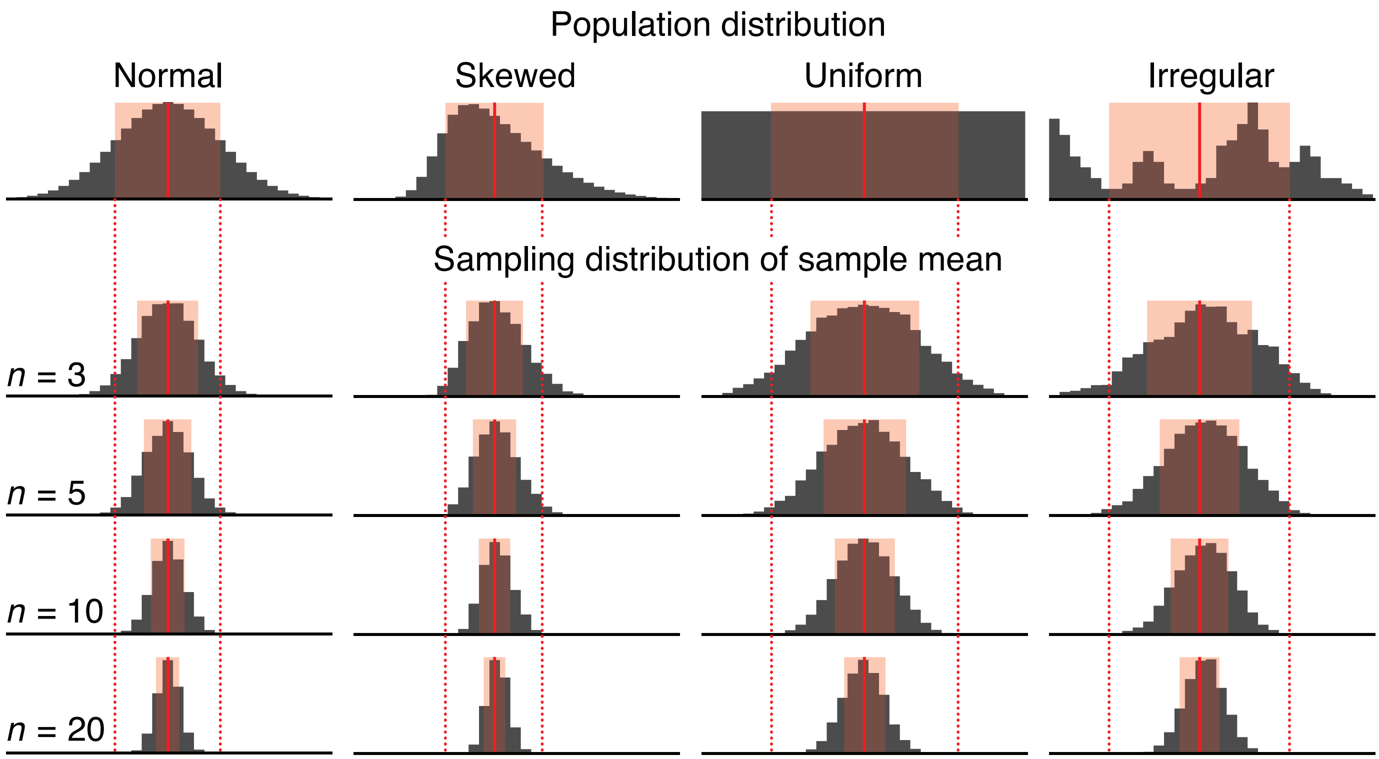
  $\rightarrow \quad E\{\sum_{i=1}^{n} y_i\} = n\pi, \ Var\{\sum_{i=1}^{n} y_i\} = n\pi(1-\pi)$

  $\rightarrow \ p = \bar{Y} \overset{approximately}{\sim} \mathcal{N}\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$

Vocabulary

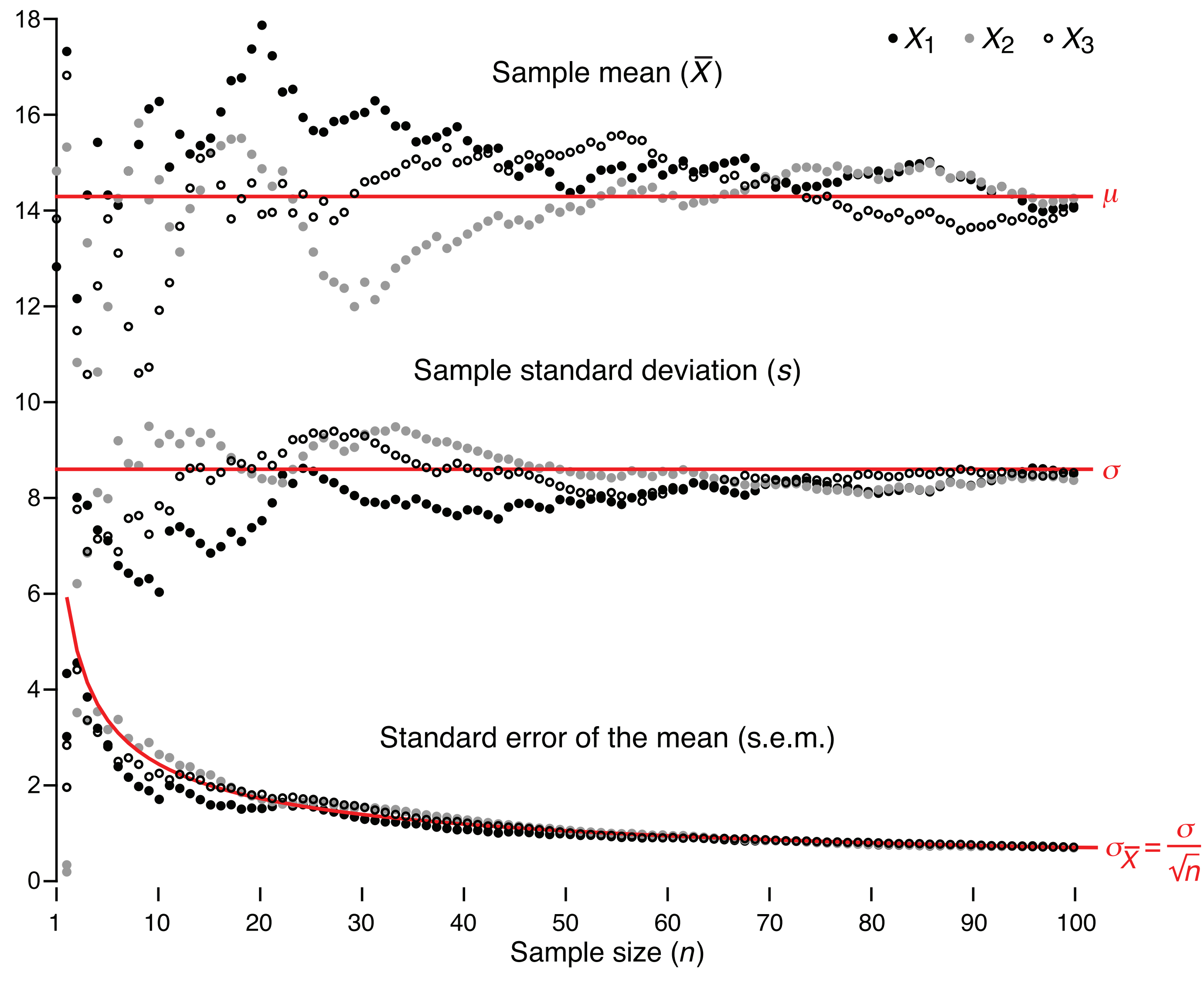
- The estimate of variation of the sampling distribution is called **standard error**
  - If Y is continuous, and $\bar{Y} \overset{approximately}{\sim} \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, then $\frac{\sigma}{\sqrt{n}}$ is the standard error of the mean
  - If Y is binary, and $\bar{Y} \overset{approximately}{\sim} \mathcal{N}\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$, then $\sqrt{\frac{\pi(1-\pi)}{n}}$ is the standard error of the sample proportion

Examples



Nature Methods, 'Points of Significance' series

Examples



Nature Methods, 'Points of Significance' series