Experiment, Sample Space, and Event

Experiment: the process of obtaining observations.

Sample space: all possible outcomes of an experiment.

Event: certain outcomes of an experiment.

Toy example 1: Coin flips.
Experiment: Flip a coin twice.
Sample space: $\{hh, ht, th, tt\}$.

Event	description
$\{hh\}$	two heads
$\{hh, ht, th\}$	at least one head
$\{hh, tt\}$	two of same side

Toy example 2: Rolls of dice.
Experiment: Roll a pair of dice.
Sample space: $\{11, 12, \ldots, 66\}$.

Event	description
$\{66\}$	double sixes
$\{12, 21\}$	total 3
$\{22, 24, \ldots, 66\}$	both even

Event Operations

Sample space is usually denoted by S, events by A, B, C, $\ldots$.

Union $A \cup B$: at least one; either; "or".

Intersection $A \cap B$: both; "and".

Complement $\bar{A}$ or A': anything but; "not".

Toy example 1: Coin flips.
$A = \{hh\}$, $B = \{hh, ht, th\}$,
$C = \{hh, tt\}$, $D = \{ht, tt\}$.

$$A \cap B = A, \quad A \cup B = B,$$
$$C \cup D = \{hh, ht, tt\},$$
$$C \cap D = \{tt\} = \bar{B},$$
$$\bar{A} = \{ht, th, tt\}.$$

- A is a subset of B: $A \subset B$.
- A and D disjoint: $A \cap D = \{\} = \Phi$.

For arbitrary A, $\Phi \subseteq A \subseteq S$.

de Morgan's Law: $\overline{A \cup B} = \bar{A} \cap \bar{B}$

$$\overline{A \cap B} = \bar{A} \cup \bar{B}$$

Trivia: $A \cup \bar{A} = S$, $A \cap \bar{A} = \Phi$.

Probability as Relative Frequency

Urn Model: Consider an urn containing n balls, of which s are black. The probability of getting a black ball in a random draw is $p = s/n$.

- Random draw: All balls have equal chance to be drawn.
- Need counting rules for computing n and s.

Toy example 1: Coin flips.
Assume a fair coin.

$$P(\{hh\}) = 1/4, \quad P(\{hh, ht, th\}) = 3/4, \quad P(\{hh, tt\}) = 2/4.$$

Toy example 2: Rolls of dice.
Assume a pair of fair dice.

$$P(\text{double sixes}) = 1/36, \quad P(\text{total } 3) = 2/36, \quad P(\text{both even}) = 9/36.$$

Objective and Subjective Probability

After rolling a die 1000 times, Alan tallied 150 sixes. He concluded that the probability of getting a six with the die is about $p = .15$.

Without knowing Alan's results, Andy was asked to assess the probability of getting a six with the die. He suggested $p = .1667$.

Who's right?

Objective: Relative frequency in repeated experiments.

Subjective: Beliefs, bets, odds, etc.

- In science, one usually is concerned about objective probability.

Alan is objective, but are 1000 rolls enough to establish "trend"?

Assuming a fair die, Andy is objective. But the assumption of fairness can be subjective.

Axioms of Probability

1. For any event A, $0 \leq P(A) \leq 1$.
2. $P(S) = 1$.
3. If $A \cap B = \Phi$, then $P(A \cup B) = P(A) + P(B)$.

- It is impossible to have $P(A) = .5$, $P(B) = .8$, and $P(A \cap B) = .1$.
- It is impossible to have $P(A) = .3$ and $P(A \cap B) = .35$.
- It is impossible to have $P(A) = .3$ and $P(\bar{A}) = .6$.

Additive Laws of Probability

1. For $A_1, \ldots, A_n$ disjoint, $P(A_1 \cup \cdots \cup A_n) = P(A_1) + \cdots P(A_n)$.
2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
3. $P(\bar{A}) = 1 - P(A)$.

A circuit chip may have etching defect (12%), crack defect (29%), and maybe both (7%). What proportion is defect-free?

$P(A) = .12$, $P(B) = .29$,
$P(A \cap B) = .07$.

$P(A \cup B) = .12 + .29 - .07 = .34$

$P(\overline{A \cup B}) = 1 - .34 = \boxed{.66}$.

Consider drawing a card from a deck of 52.

$$P(\text{Red K}) = \frac{2}{52}$$
$$P(\{3,4,5,6\}) = \frac{4}{13}$$
$$P(\text{Red A or Blk Q}) = \frac{4}{52}$$
$$P(\text{Red or K}) = \frac{1}{2} + \frac{1}{13} - \frac{2}{52}$$

Counting Rule: Multiplication

Consider k drawers containing n_1, n_2, ..., n_k distinctive items, respectively. The total number of ways to pick one item each from the k drawers is: $\boxed{n_1 n_2 \cdots n_k}$.

- The number of outcomes from k coin flips: $\boxed{2^k}$.
- The number of outcomes from k rolls of a die: $\boxed{6^k}$.
- The number of even-even outcomes from 2 rolls of a die: $\boxed{3^2}$.
- The number of letter-digit combinations: $\boxed{26(10)}$.
- The probability of winning pick-3 lottery: $\boxed{1/10^3}$.

Counting Rule: Permutation

Consider a pool of n candidates. The number of ways to rank the top r is: $\boxed{n(n-1)\cdots(n-r+1) = \frac{n!}{(n-r)!} = P_r^n = P_{r,n}}$.

- The number of ways to assign the top 3 seeds in the Big 10 conference: $\boxed{P_3^{14} = 14(13)(12) = 2184}$.
- The number of ways to assign 4 brands of grass seeds to 4 plots of experimental field: $\boxed{P_4^4 = 4! = 24}$.
- The number of ways to seat Wheel-of-Fortune players: $\boxed{3! = 6}$.

In R, $n!$ can be calculated via `factorial(n)`.

Counting Rule: Combination

Consider a pool of n candidates. The number of ways to pick r from the pool is: $\boxed{\frac{P_r^n}{r!} = \frac{n!}{r!(n-r)!} = \binom{n}{r} = C_r^n = C_{r,n} = C_{n-r}^n}$.

- The number of possible starting lineups for each basketball team: $\boxed{C_5^{12} = 792}$.
- The probability of hitting a jack-pot: $\boxed{\frac{1}{C_6^{44}} = \frac{1}{7059052}}$.
- The probability of winning a Power-Ball: $\boxed{\frac{1}{44C_5^{44}} = \frac{1}{47784352}}$.

In R, C_r^n can be calculated via `choose(n,r)`.

Counting Rules: Examples

Excercise 2.90 on page 88: Pick a crew of 3 from 20 machinists.

1. There are $C_3^{20} = 1140$ possible crews.
2. $C_0^1 C_3^{19} = 969$ crews do not include the best machinist.
3. $C_1^5 C_2^{15} + C_2^5 C_1^{15} + C_3^5 C_0^{15} = 685$ crews have at least one of the 5 best machinists.

Excercise 2.40 on page 72: 3 each of A, B, C, D are to be arranged into a chain.

1. There are $P_{12}^{12}/(P_3^3)^4 = 12!/(3!)^4 = 369600$ possible chains.
2. There are $P_4^4 = 4! = 24$ chains with the same letter next to each other.

Conditional Probability and Independence

Conditional Probability: $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$.

Independence: A and B are independent if $P(A|B) = P(A)$.

Chip defects.

$P(A) = .12$, $P(B) = .29$,

$P(A \cap B) = .07$.

$$P(A|B) = \frac{.07}{.29} = .241 \neq .12$$

$$P(B|A) = \frac{.07}{.12} = .583 \neq .29$$

A and B are dependent.

- Conditioning reduces sample space.
- Conditioning can simplify calculation.
- Independence is often established before calculation.

Multiplicative Laws of Probability

1. $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$.
2. For A and B independent, $P(A \cap B) = P(A)P(B)$.

Consider drawing two cards w/o replacement from a deck of 52.

$$P(\text{Both Red}) = \frac{26}{52}\frac{25}{51}$$

$$P(\text{Two Colors}) = 2\frac{26}{52}\frac{26}{51}$$

Using counting rules,

$$P(\text{Both Red}) = \frac{C_2^{26}}{C_2^{52}}$$

$$P(\text{Two Colors}) = \frac{C_1^{26}C_1^{26}}{C_2^{52}}$$

Consider five rolls of a fair die.

$$P(\bar{6}\bar{6}\bar{6}\bar{6}\bar{6}) = \left(\frac{5}{6}\right)^5$$

$$P(\bar{6}\bar{6}\bar{6}\bar{6}6) = \left(\frac{5}{6}\right)^4\left(\frac{1}{6}\right)$$

$$P(o55aa) = \left(\frac{3}{6}\right)\left(\frac{1}{6}\right)^2$$

$$P(12345) = \left(\frac{1}{6}\right)^5$$

where o is odd and a is any.

Bayes' Theorem

Let B_i be a partition of S: $S = B_1 \cup \cdots \cup B_k$, B_i disjoint.

$$\boxed{P(B_j|A) = \frac{P(B_j)P(A|B_j)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)}}.$$

Consider an AIDS test.

$$P(T^+|D) = .98,$$

$$P(T^-|\bar{D}) = .99.$$

	D	$\bar{D}$
T^+	.000019(.98)	.999981(.01)
T^-	.000019(.02)	.999981(.99)

Also known,

$$P(D) = .000019.$$

Find $P(\bar{D}|T^+)$, $P(D|T^-)$.

$$P(\bar{D}|T^+) = \frac{.999981(.01)}{.01001843} = \boxed{.998},$$

$$P(D|T^-) = \frac{.000019(.02)}{.9899816} = \boxed{3.84 \times 10^{-7}}.$$

Random Variables

Random variable assumes numerical values according to the outcome of an experiment. Random variables are usually denoted by X, Y, Z, etc.

Examples of random variables:

1. No. of heads in 5 coin flips.
2. Total of a pair of dice.
3. No. of crossovers between a pair of chromosomes.
4. Miles between oil changes.
5. Temperature at noon.

- Continuous r.v. takes value on a continuum such as $[0, 1]$; discrete r.v. takes value on a discrete set such as $\{1, 2, \ldots\}$.
- A r.v. is *not* a fixed number.
- The behavior of a r.v. is characterized by its **distribution**.

Discrete Probability Distribution

Discrete probability distribution consists of a set of possible values and the associated probabilities:

x	x_1	x_2	$\cdots$
$P(x)$	p_1	p_2	$\cdots$

$X =$ no. of heads in 3 coin flips.

x	0	1	2	3
$P(x)$	1/8	3/8	3/8	1/8

$P(X = 2) = P(2) = 3/8$.

- The set $\{x_1, x_2, \dots\}$ can be finite or infinite, and x_i's need not to be integers.
- $P(X = x_i) = P(x_i) = p_i$.
- $0 \leq p_i \leq 1$. $\sum_i p_i = 1$.

Mean and Standard Deviation

Mean summarizes the central location: $\boxed{\mu = \sum_i x_i p_i}$.

Standard deviation summarizes the spread: $\sigma = \sqrt{\sigma^2}$, with σ^2 the variance given by $\boxed{\sigma^2 = \sum_i (x_i - \mu)^2 p_i = \sum_i x_i^2 p_i - \mu^2}$.

X = no. of heads in 3 coin flips.

x	0	1	2	3
$P(x)$	1/8	3/8	3/8	1/8

$$\mu = 0\frac{1}{8} + 1\frac{3}{8} + 2\frac{3}{8} + 3\frac{1}{8} = \boxed{1.5},$$

$$\sigma^2 = \frac{3}{8} + \frac{12}{8} + \frac{9}{8} - 1.5^2 = \boxed{.75}.$$

- On average, one gets 1.5 heads in 3 flips of a fair coin.
- Imagine a data set of size 8m, with 1m 0's, 3m 1's, 3m 2's, and 1m 3's. One should get $\bar{x} = 1.5$ and $s^2 = .75$.
- The mean is also called the **expected value**.

Bernoulli Trials

A series of trials are **Bernoulli Trials** if:

1. The outcome of each trial is binary, say Y/N.
2. $P(Y) = p$ remains the same for all trials.
3. The trials are independent.

Examples of Bernoulli trials:

- Multiple flips of a coin.
- Defectiveness of light bulbs on the store shelves.
- Genders of people picked from the phone books.
- Responses of patients to a certain antibiotic.

Binomial and Geometric Distributions

Binomial distribution characterizes the total number of Y's in a fixed number of Bernoulli trials: $\text{Bin}(n, p)$.

Possible values	$x = 0, 1, 2, \ldots, n$
Probabilities	$P(x) = C_x^n p^x (1-p)^{n-x}$
Parameters	$\mu = np$, $\sigma = \sqrt{np(1-p)}$

Geometric distribution characterizes the trial number of the first Y in a series of Bernoulli trials.

Possible values	$x = 1, 2, \ldots$
Probabilities	$P(x) = (1-p)^{x-1} p$
Parameters	$\mu = 1/p$, $\sigma = \sqrt{1-p}/p$

Binomial Probabilities

When two carriers of gene for albinism marry, each child has 1/4 chance of being albino. Let X be the number of albino children in such a family with 5 children. $X \sim \text{Bin}(5, .25)$.

$$P(0) = C_0^5(.25)^0(.75)^5 = .237$$

$$P(1) = C_1^5(.25)(.75)^4 = .396$$

$$P(X \leq 1) = .237 + .396 = .633$$

$$P(X \geq 1) = 1 - .237 = .763$$

About one-third Americans 20 or older are at high risk for coronary disease due to high cholesterol levels. Let X be the number of high risk adults in a sample of 20. $X \sim \text{Bin}(20, 1/3)$.

$$P(3) = C_3^{20}(\frac{1}{3})^3(\frac{2}{3})^{17} = .0429$$

$$P(0) = (\frac{2}{3})^{20} = .0003$$

$$P(X \geq 1) = 1 - .0003 = .9997$$

- $b(x; n, p)$ can be calculated in R via `dbinom(x,n,p)`.

Binomial Probabilities

The sex-ratio data of 72069 six-child families are given below.

Boys	Freq.	Prop.	Fit
0	1096	.0152	.0130
1	6233	.0865	.0830
2	15700	.2178	.2202
3	22221	.3083	.3117
4	17332	.2405	.2481
5	7908	.1097	.1053
6	1579	.0219	.0186

Are children within families independent?

- The overall proportion of boys is $\hat{p} = .5149$, which corresponds to a 106:100 sex-ratio.
- The fit column contains probabilities of $\text{Bin}(6, \hat{p})$.
- The data seem to favor "runs" within families. We will come back later with formal analysis.

If $X \sim \text{Bin}(n, p)$, $Y \sim \text{Bin}(m, p)$, and independent, then

$$X + Y \sim \text{Bin}(n + m, p).$$

Cumulative Distribution Function

Cumulative distribution function is often tabulated to facilitate probability calculations: $F(x) = \sum_{x_i \le x} P(x_i)$.

Consider X the number of heads in 6 flips of a fair coin.

$$P(X=3) = P(3) = .3125$$

$$P(X=3) = F(3) - F(2) = .3124$$

$$P(X \ge 2) = 1 - F(1) = .8906$$

$$P(X \le 3) = F(3) = .6562$$

- It is easier to calculate probabilities of "blocks" using $F(x)$.

Bin(6,.5) distribution:

x	$P(x)$	$F(x)$
0	.0156	.0156
1	.0938	.1094
2	.2344	.3438
3	.3125	.6562
4	.2344	.8906
5	.0938	.9844
6	.0156	1.0000

Hypergeometric Distribution

Hypergeometric distribution characterizes draws from an urn *without replacement*: N balls in the urn with k of them black; draw n balls from the urn and count X blacks.

Possible values	$x = \max(0, n-(N-k)), \ldots, \min(n, k)$
Probabilities	$P(x) = \frac{C_x^k C_{n-x}^{N-k}}{C_n^N}$
Parameters	$\mu = n\frac{k}{N}$, $\sigma = \sqrt{n\frac{k}{N}(1-\frac{k}{N})\frac{N-n}{N-1}}$

Of 8 patients given a cold medicine 5 recovered in 3 days. Of 10 given placebo 4 did so. What's the probability that this just happened "by chance"? $C_5^8 C_4^{10}/C_9^{18}$

- With replacement one gets Binomial with $p = k/N$.
- The same μ as for Binomial, but smaller σ.

Poisson Distribution

Poisson distribution characterizes the number of incidences occurring in a time interval or space volume: Poisson(λ).

Possible values	$x = 0, 1, 2, \ldots$
Probabilities	$P(x) = \lambda^x e^{-\lambda}/x!$
Parameters	$\mu = \lambda$, $\sigma = \sqrt{\lambda}$

Examples of Poisson r.v.:

1. No. of crossovers on genome.
2. No. of bacteria in fluid.
3. No. of arrivals at counter.
4. No. of dial-ins at ISP.

- Need to specify volume.
- λ determines intensity.

Poisson Process: cut T into tiny pieces Δt.

1. P(a hit in Δt) $\propto$ size.
2. $P(2_+$ hits in $\Delta t) \approx 0$.
3. Pieces are independent.

Poisson Probabilities

A brand of wall paper averages $.3/ft^2$ coating blemishes and $.1/ft^2$ printing blemishes. The two kinds are independent.

$$P(3 \text{ cb's on } 10ft^2) = 3^3 e^{-3}/3!$$

$$P(4 \text{ pb's on } 20ft^2) = 2^4 e^{-2}/4!$$

$$P(\text{None on } 7ft^2) = 2.8^0 e^{-2.8}/0!$$

If $X \sim \text{Pois}(\lambda_1)$, $Y \sim \text{Pois}(\lambda_2)$, and independent, then

$$X + Y \sim \text{Pois}(\lambda_1 + \lambda_2).$$

The number of crossovers X between two points on a genome roughly follows a Poisson distribution. When X is odd, one has a recombination.

$$P(X \text{ odd}) = \sum_{i=0}^{\infty} \frac{\lambda^{2i+1} e^{-\lambda}}{(2i+1)!} = \frac{1}{2}(1 - e^{-2\lambda})$$

Distance on genome is measured in terms of λ, with $\lambda = 1$ roughly a Morgan (M) and $\lambda = .01$ a cM.

Probability Approximations

Binomial approximation of hypergeometric: When N is large and n/N is small, say $n/N \leq .05$,

$$C_x^k C_{n-x}^{N-k} / C_n^N \approx C_x^n p^x (1-p)^{n-x}, \text{ where } p = k/N.$$

Poisson approximation of Binomial: When n is large and $np \leq 5$, say, $C_x^n p^x (1-p)^{n-x} \approx \lambda^x e^{-\lambda}/x!$, where $\lambda = np$.

With $N = 100$, $k = 10$, $n = 5$, and $x = 2$,

$$\frac{C_2^{10} C_3^{90}}{C_5^{100}} = .0702$$

$$C_2^5 (.1)^2 (.9)^3 = .0729$$

With $n = 100$, $p = .03$, and $x = 2$,

$$C_2^{100} (.03)^2 (.97)^{98} = .2252$$

$$3^2 e^{-3}/2! = .2240$$

For n large and p "fixed", use **normal approx.** discussed later.

Probability Distributions in R

R provides four utility functions for each of the many commonly used distributions: `r-` for data simulation, `d-` for probability density function (pdf), `p-` for cumulative distribution function (cdf), and `q-` for quantiles (inverse of cdf).

```
rbinom(7,5,0.6); rpois(10,5.5); rhyper(7,9,6,5)
dbinom(0:5,5,0.6); dpois(0:10,5.5); dhyper(0:5,9,6,5)
pbinom(0:5,5,0.6); cumsum(dbinom(0:5,5,0.6))
qpois(c(0,.25,.5,.75,1),5.5); ppois(0:10,5.5)
```

Table A.1 lists the results of `pbinom(x,n,p)` for selected (n, p).
Table A.2 lists the results of `ppois(x,lambda)` for selected λ.

Continuous Probability Distribution

Probability density function (pdf) specifies a continuous probability distribution: $f(x) \geq 0$.

Cumulative distribution function (cdf) facilitates the calculation: $F(x) = \int_{-\infty}^{x} f(t)dt = P(X \leq x) = P(-\infty, x]$.

"Throw darts" on $(0,1)$ and record the locations X: $U(0,1)$.

$$f(x) = \begin{cases} 0, & x \leq 0, \\ 1, & 0 < x < 1, \\ 0, & 1 \leq x. \end{cases}$$

$$F(x) = \begin{cases} 0, & x \leq 0, \\ x, & 0 < x < 1, \\ 1, & 1 \leq x. \end{cases}$$

- The pdf $f(x) \geq 0$ is nonnegative.
- The total area under $f(x)$ is 1.
- $P(a,b] = F(b) - F(a)$: the area under $f(x)$ between a and b.
- $P\{a\} = 0$, so $P[a,b] = P(a,b]$.
- $F(-\infty) = 0$, $F(\infty) = 1$, $F(x) \uparrow$.
- If $F(a) = p$, then a is called the $100p$-th percentile.

Mean and Standard Deviation

Mean summarizes the central location: $\mu = \int_{-\infty}^{\infty} x f(x) dx$.

Standard deviation summarizes the spread: $\sigma = \sqrt{\sigma^2}$, with the variance $\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2$.

Darts on $(0,1)$: $U(0,1)$

$$\mu = \int_0^1 x dx = \frac{1}{2}$$

$$\sigma^2 = \int_0^1 (x - \frac{1}{2})^2 dx$$

$$= \int_0^1 x^2 dx - (\frac{1}{2})^2 = \frac{1}{12}$$

- The average position is at the center.
- Imagine throwing millions of darts on $(0,1)$ and recording locations x_i. One should get $\bar{x} = \frac{1}{2}$ and $s^2 = \frac{1}{12}$.

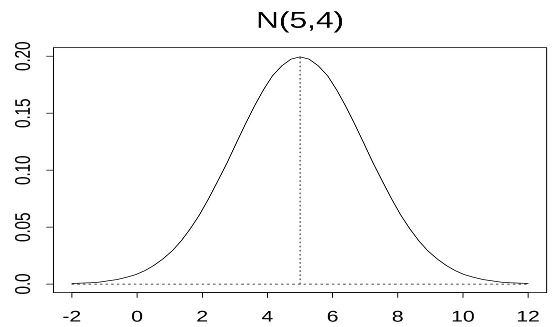
Normal Distribution

Normal distribution is the default "measurement error" model and is the "attraction point" under averaging: $N(\mu, \sigma^2)$.

"Measurement errors":

1. Heights of Purdue students.
2. Miles between oil changes.
3. Yields per acre.
4. Weights of tuna cans.
5. SAT scores of students.
6. Reported particle counts.

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



- $f(x)$ is symmetric w.r.t. μ.
- $F(x)$ has no explicit form.

Standard Normal Distribution

Normality specifies the shape, and is preserved under linear transformation: If $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$, then $Y \sim N(a\mu + b, a^2\sigma^2)$.

Standard normal distribution: $N(0, 1)$.

Standardization converts an arbitrary normal r.v. to a standard normal r.v.: If $X \sim N(\mu, \sigma^2)$, then $(X - \mu)/\sigma \sim N(0, 1)$.

The heights of male students on a university campus follow $N(68, 25)$, in *in*, or $N(2.54(68), (2.54)^2 25)$, in *cm*.

- To make calculations concerning normal r.v., it suffices to know $F(x)$ for $N(0, 1)$.
- $F(x)$ for $N(0, 1)$ is finely tabulated.

Standard Normal Probabilities

- By convention, one writes $Z \sim N(0,1)$ and $P(Z \le z) = \Phi(z)$.
- $\Phi(z)$ is tabulated (Table A.3 inside the front cover of textbook).

Using $\Phi(z)$ table and symmetry,

$$\begin{aligned}
P(Z \le 0) &= P(Z \ge 0) = .5, \\
P(Z \le .55) &= \Phi(.55) = .7088, \\
P(Z > .33) &= 1 - \Phi(.33) = .3707, \\
P(Z \le -2) &= \Phi(-2) = .0228, \\
P[1.2, 2] &= \Phi(2) - \Phi(1.2) = .0923, \\
P(-1.2, 2) &= \Phi(2) - \Phi(-1.2) \\
&= .8621, \\
P[-1, 1] &= \Phi(1) - \Phi(-1) = .6826.
\end{aligned}$$

Find a, b, c, given

$$P(0, a) = \Phi(a) - \Phi(0) = .4,$$
$$P(-2, b) = \Phi(b) - \Phi(-2) = .3,$$
$$P[c, 1.5] = \Phi(1.5) - \Phi(c) = .6.$$

As $\Phi(a) = .9$, so $\boxed{a \approx 1.28}$.

As $\Phi(b) = .3228$, or $\Phi(-b) = .6772$, so $\boxed{b = -.46}$.

As $\Phi(c) = .3332$, or $\Phi(-c) = .6668$, so $\boxed{c \approx -.43}$.

Normal Probabilities

For $X \sim N(\mu, \sigma^2)$, standardize via $Z = \frac{X-\mu}{\sigma}$ and use $\Phi(z)$.

$$P(a \le X \le b) = P(\tfrac{a-\mu}{\sigma} \le \tfrac{X-\mu}{\sigma} \le \tfrac{b-\mu}{\sigma}) = \Phi(\tfrac{b-\mu}{\sigma}) - \Phi(\tfrac{a-\mu}{\sigma})$$

- Key to calculations: *Standardize, standardize, standardize.*

Consider $X \sim N(68, 25)$.

$P(X \le 68) = P(X \ge 68) = .5$,

$P(X > 72) = 1 - \Phi(.8) = .2119$,

$P(63, 73) = \Phi(1) - \Phi(-1) = .6826$,

Find a such that $P(60, a) = .6$.

Since $\Phi(\frac{a-68}{5}) - \Phi(-1.6) = .6$, or $\Phi(\frac{a-68}{5}) = .6548$, one has $\frac{a-68}{5} \approx .4$, or $\boxed{a \approx 70}$.

Find μ, given $X \sim N(\mu, 36)$ and $P(X < 5) = .4$.

Since $\Phi(\frac{5-\mu}{6}) = .4$, $-\frac{5-\mu}{6} \approx .25$, or $\boxed{\mu \approx 6.5}$.

Find σ, given $X \sim N(6, \sigma^2)$ and $P(X < 5) = .4$.

Since $\Phi(\frac{5-6}{\sigma}) = .4$, $\frac{1}{\sigma} \approx .25$, or $\boxed{\sigma \approx 4}$.

Normal Percentiles

- By convention, z_α denotes the $100(1-\alpha)$th percentile of $N(0,1)$: $\Phi(z_\alpha) = 1-\alpha$, or $P(Z > z_\alpha) = \alpha$.
- For $X \sim N(\mu, \sigma^2)$, $P(|X-\mu| \le k\sigma) = P(|Z| \le k) = 2\Phi(k) - 1$.

α	z_α
.005	2.576
.010	2.326
.025	1.960
.050	1.645

k	$P(\|Z\| \le k)$
1	.6827
2	.9545
3	.9973

A builder believes the material cost for a new project will behave as $N(60,16)$ and the labor cost as $N(30,9)$. What is his chance to keep the total cost below 100?

The total cost behaves as $N(90,25)$, so

$$P(0,100) = \Phi(2) - \Phi(-\infty) = .9772.$$

If $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, and independent, then $X+Y \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$.

Exponential Distribution

Exponential distribution provides a simple model for life time data: $f(x) = \lambda e^{-\lambda x}$, $x > 0$.

- $F(x) = 1 - e^{-\lambda x}$, $x > 0$; $\mu = \frac{1}{\lambda}$, $\sigma = \frac{1}{\lambda}$.
- When the number of events on unit time interval follows Poisson, the time lag between events follows exponential.

On average, a surveillance camera can run 50 days without being reset.

$$P(\text{run for } 60_{+} \text{ days}) = e^{-1.2}$$

$$P(\text{reset in 20 days}) = 1 - e^{-.4}$$

On average, a customer arrives at a bank counter every 4 minutes.

$$P(3 \text{ within } 6 \text{ min}) = \frac{1.5^3 e^{-1.5}}{3!}$$

$$P(\text{next in 3 min}) = 1 - e^{-3/4}$$

Probability Distributions in R

The R utility functions for normal and exponential distributions have default parameter values that may be overrided.

```
pnorm(.55); qnorm(.7088)
pnorm(72,mean=68,sd=5); pnorm((72-68)/5)
qnorm(.6548,68,5); 5*qnorm(.6548)+68
pnorm(1)-pnorm(-1); qnorm(1-.025)

pexp(60,rate=1/50); pexp(60/50)
qexp(.5,1/50); 50*qexp(.5)
```

Table A.3 lists the results of `pnorm(z,mean=0,sd=1)`.

Simple Random Sampling

Simple random sampling selects with equal chance from (available) members of population. The resulting sample is a **simple random sample**.

Consider an urn containing N balls with numbers x_i written on them. Draw n balls from the urn.

- Equal chance for C_n^N possible samples w/o replacement: *finite population.*
- Equal chance for N^n possible samples w/ replacement: *infinite population.*

- The x_i's need not to be all different.
- One can let $N \to \infty$, but then can not sample w/o replacement.
- A sample from an infinite population consists of *independent, identically distributed (i.i.d.)* observations.

Sampling Distribution

Sampling distribution describe the behavior of **sample statistics** such as $\bar{x}$ and s^2.

In a certain human population, 30% of the individuals have "superior" distance vision (20/15 or better). Consider the sample proportion of superior vision,

$$\hat{p} = X/n,$$

where X is the number of people in the sample with superior vision. Find the sampling distribution of $\hat{p}$ for $n = 20$.

Clearly, $X = 20\hat{p} \sim \text{Bin}(20, .3)$.

Possible values for $\hat{p}$ are

$$\{0,.05, .10, \ldots, .95, 1\},$$

with the probabilities given by

$$P(\hat{p}{=}x) = C_{20x}^{20}(.3)^{20x}(.7)^{20(1-x)}.$$

For example,

$$P(\hat{p}{=}.3) = C_{6}^{20}(.3)^{6}(.7)^{14} = .192.$$

In R, use `dbinom(0:20,20,.3)`.

Simulating Sampling Distributions

When analytical derivation is cumbersome or infeasible, one may use simulation to obtain the sampling distribution.

Example: The sample median of 17 r.v.'s from $N(0,1)$.

```
x <- matrix(rnorm(170000),17,10000)
md <- apply(x,2,median)
hist(md,nclass=50); plot(density(md))
```

Example: The largest of 5 Poisson counts from Poisson(3.3).

```
x <- matrix(rpois(50000,3.3),5,10000)
mx <- apply(x,2,max)
table(mx); table(mx)/10000
ppois(1:13,3.3)^5-ppois(0:12,3.3)^5
```

Sampling Distribution of $\bar{X}$

Use upper case $\bar{X}$ to denote the sample mean as a r.v. For an infinite population with mean μ and standard deviation σ,

$$\mu_{\bar{X}} = \mu \quad \text{and} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

The heights of male students on a large university campus have $\mu = 68$ and $\sigma = 5$. Consider $\bar{X}$ with sample size $n = 25$.

$$\mu_{\bar{X}} = \mu = 68$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{25}} = 1$$

- $\bar{X}$ is more concentrated around μ.
- To double the "accuracy" of $\bar{X}$, one needs to quadruple the sample size n.

For finite population, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$.

Central Limit Theorem

Consider an infinite population with mean μ and standard deviation σ. For n large,

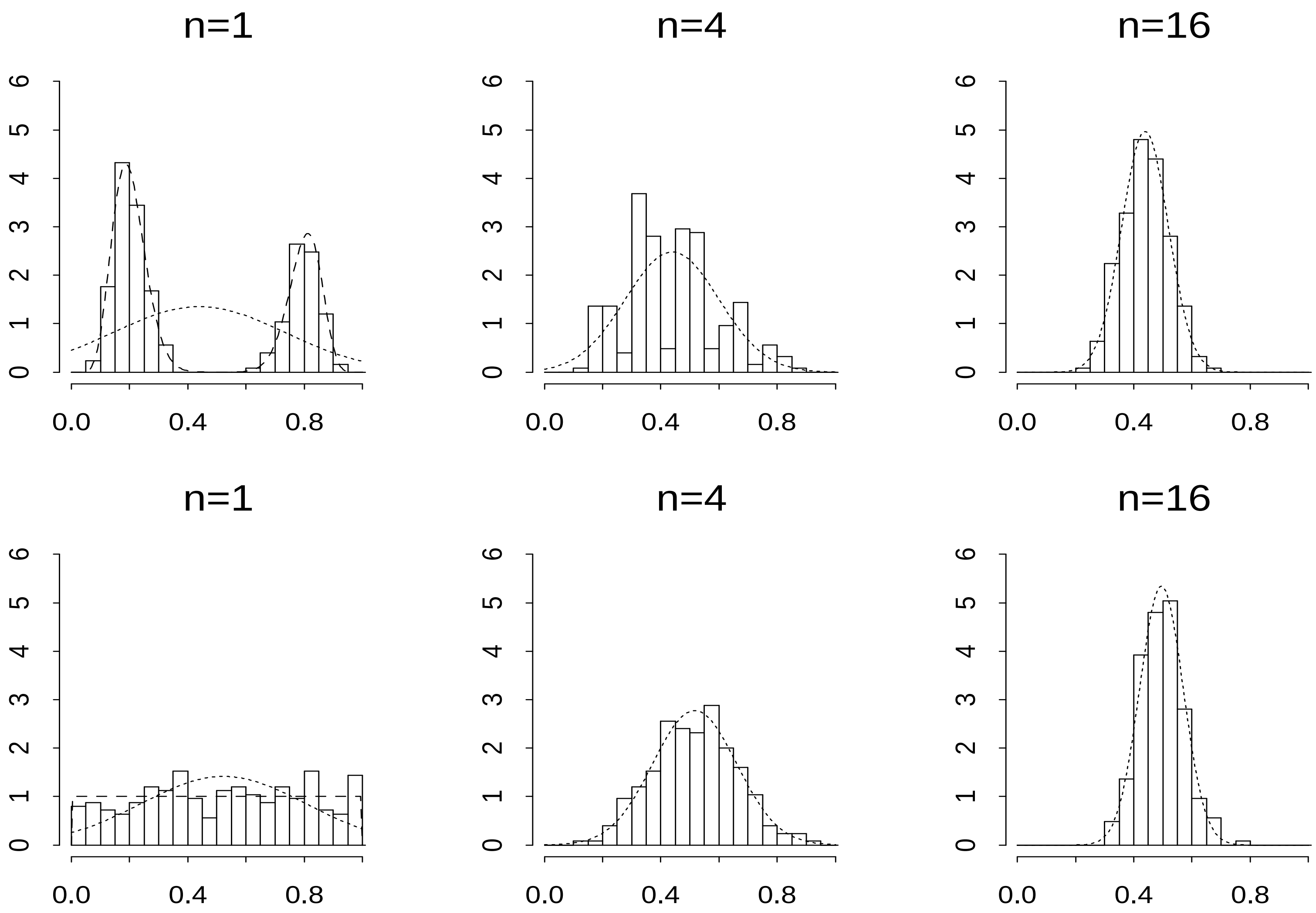
$$P(\frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}} \leq z) \approx \Phi(z).$$

Given $\mu = 68$ and $\sigma = 5$. Find the probability for the average height of $n = 25$ students to exceed 70.

$$P(\bar{X} > 70) = P(\frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}} > \frac{70-68}{1})$$
$$\approx 1 - \Phi(2) = \boxed{.0228}.$$

- The shape of sampling distribution approaches normal as $n \to \infty$. Usually, an $n \geq 30$ is sufficiently large for CLT to kick in.
- For normal population, $\bar{X}$ is always normal, regardless of n.

Effects of Sample Size



Normal Approximation of Binomial

Recall that if $X \sim \text{Bin}(n,p)$, then $X = \sum_{i=1}^n X_i$, where $X_i \sim \text{Bin}(1,p)$. By the Central Limit Theorem,

$$P\left(\frac{X-np}{\sqrt{np(1-p)}} \le z\right) = P\left(\frac{X/n-p}{\sqrt{p(1-p)/n}} \le z\right) \approx \Phi(z)$$

Consider $X \sim \text{Bin}(25, .3)$.

$P(X \le 8) = .6769$

$$\begin{aligned} P(X \le 8) &= P(X \le 8.5) \\ &\approx \Phi\left(\frac{8.5-7.5}{2.291}\right) = .6687 \end{aligned}$$

$P(X = 8) = .1651$

$$\begin{aligned} P(X = 8) &= P(7.5 \le X \le 8.5) \\ &\approx \Phi(.436) - \Phi(0) = .1687 \end{aligned}$$

- When a continuous distribution is used to approximate a discrete one, **continuity correction** is needed to preserve accuracy.
- For $np, n(1-p) \ge 5$, the approximation is reasonably accurate.

Summary

Probability provides a mathematical language for the study of chance phenomena. One can use probability to describe the anticipated behavior of samples from populations with known characteristics.

Topics covered:

- **Events** in a **sample space** describe outcomes of a (stochastic) **experiment**.
- Probability is defined **objectively** as the relative frequency of events or **subjectively** as odds or through bets or beliefs. One needs **counting rules** for the former.
- Probabilities of related events have to conform to certain **axioms** and follow certain **laws**.

- Perceived chance, or probability, of an event may depend on what one knows about the occurrence of other events. There lies the joy of **independence** and **conditional probability**.
- **Random variables** assume numerical values according to outcomes of stochastic experiments. **Probability distributions** describe the behavior of random variables.
- Common distributions: **Binomial**, **Poisson**, **Normal**.
- **Sampling distribution** describes the behavior of sample **statistics**. **Central Limit Theorem** lays technical foundation for basic statistical inferences.