Outline

Example: Lake Clarity

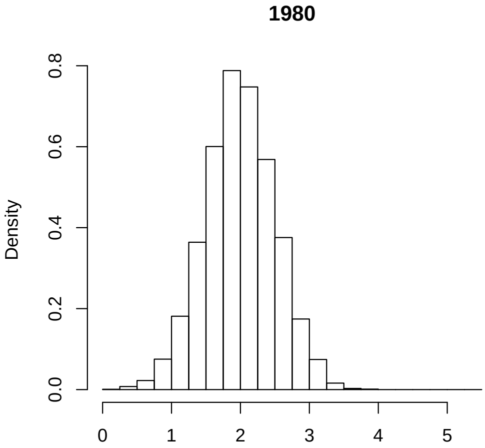
- Water clarity is an important indicator of the health of a lake.
- A measuring device called a Secchi disk provides a relatively inexpensive way of measuring lake water clarity.
- Typically used in the deepest part of the main basin of a lake, the Secchi disk is lowered into the water and the depth at which it is no longer visible is recorded.
- In an environmental monitoring program, Secchi depths were sampled repeatedly over time at many lakes.
- An objective is to determine whether the lake water clarity has changed over time.
- Our primary interest is in the results obtained from 10,000 lakes that were measured both in 1980 and in 1990.

Secchi Depth 1980 vs. 1990

lake	d80	d90
1	1.69	1.89
2	2.09	2.17
3	1.58	1.67
4	2.80	2.82
5	2.16	2.52
6	1.59	1.93
7	2.24	2.70
8	2.37	3.20
...		
9993	1.76	2.06
9994	1.99	2.46
9995	2.10	2.63
9996	2.48	3.85
9997	2.22	2.42
9998	2.25	2.99
9999	2.45	2.43
10000	2.13	2.98

Secchi Depth 1980

A histogram of the values (in meters):



- The mean is

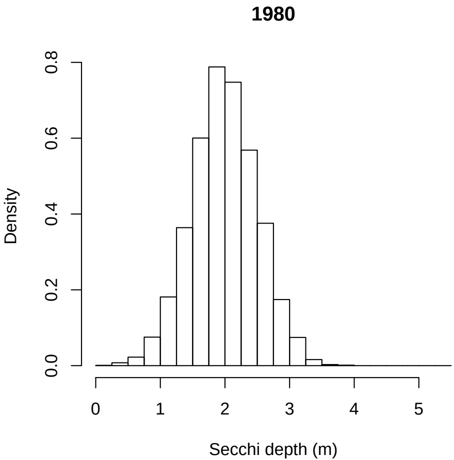
  ____

- The standard deviation is

  ____

Secchi Depth 1980 $P(Y \leq 0.25)$

- Consider **one lake** drawn **at random** from the list of 10,000 lakes.
- The probability that the Secchi depth of this lake is no greater than 0.25 m is ____________



Outline

Random Variable

- A **random variable** (r.v.) is a variable that takes its values according to a random process.
- Random variables are often denoted by capital letters, like W, X, Y, or Z.
- Lake clarity example: Let the random variable Y denote the Secchi depth value for a lake sampled at random from the 1980 population.
- We refer to some statement about a random variable as an **event**.
- Lake clarity example: The event that Y is no greater than 0.25 m, or "$Y \leq 0.25$"

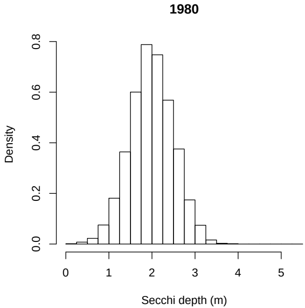
Probability of an Event

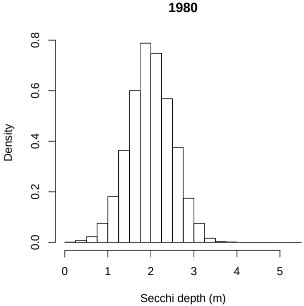
- Given a histogram of the population, and if a random variable Y represents a single random observation from the population, then probabilities of events for Y can be determined by looking at corresponding areas under the histogram.
- Lake clarity example:
  - $P(Y \leq 0.25) = 0.0002$
  - $P(Y \leq 1) = 0.0265$
  - $P(Y > 3) = 0.0236$
  - $P(1 < Y \leq 3) = 0.9499$
  - $P(1 < Y \leq 2) = 0.4834$
  - $P(1.7 < Y \leq 2.3) =??$

Upper vs. Lower Tail Probability

A possible value of a random variable is usually denoted by lower case letters, like y in $P(Y \leq y)$ or y in $P(Y > y)$.

- Lower tail (or, left tail) probability: $P(Y \leq y)$ or $P(Y < y)$.
- Upper tail (or, right tail) probability: $P(Y \geq y)$ or $P(Y > y)$.





Population Characteristics

Useful population characteristics about a random variable are:

- Population mean μ
  - A typical value of a random variable.
  - Also known as the expectation of a random variable.
  - Notation: $E(Y)$ or μ_Y.
  - Lake clarity example: $\mu_Y = E(Y) = 2.0$.
- Population standard deviation σ
  - A typical deviation of a random variable.
  - Notation: σ_Y.
  - Lake clarity example: $\sigma_Y = 0.5$.
- Population variance σ^2
  - Square of the population standard deviation.
  - Notation: $Var(Y)$ or σ_Y^2.
  - Lake clarity example: $\sigma_Y^2 = Var(Y) = 0.25$.

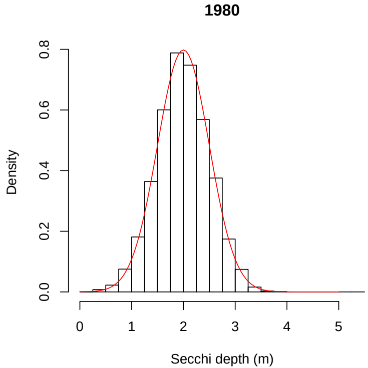
Properties of Expectation, Variance, and SD

- Let $Y^* = Y + c$.
  - $\mu_{Y^*} = E(Y^*) =$ ______________
  - $\sigma^2_{Y^*} = Var(Y^*)$ ______________
  - $\sigma_{Y^*} = \sqrt{Var(Y^*)}$ ______________
- Let $Y^* = kY$.
  - $\mu_{Y^*} = E(Y^*) =$ ______________
  - $\sigma^2_{Y^*} = Var(Y^*)$ ______________
  - $\sigma_{Y^*} = \sqrt{Var(Y^*)}$ ______________

		Population	
Random Variable	Mean	Variance	SD
Y	$\mu_Y = E(Y)$	$\sigma^2_Y = Var(Y)$	$\sigma_Y = \sqrt{Var(Y)}$
$Y + c$			
kY			

Bell-Shaped Curve

- Histograms of many populations have a similar shape and can be well-approximated by a bell-shaped curve.
- If the population can be well-approximated by a bell-shaped curve, then we can determine probabilities for the population by simply referencing the bell-shaped curve directly.

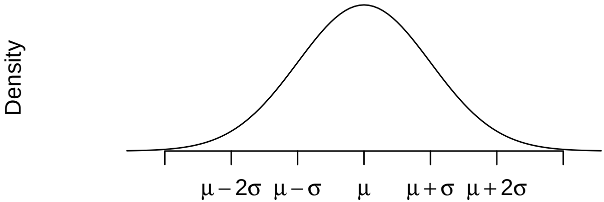


Normal Distribution

- **Normal distribution** is also known as a **Gaussian distribution.**
- The density function is given by the equation:

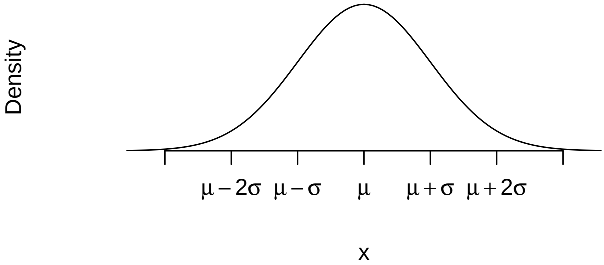
$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$

- μ and σ^2 are the **parameters** of the curve.
- $Y \sim N(\mu, \sigma^2)$ denotes that a random variable Y follows a normal distribution with mean μ and variance σ^2 (i.e. standard deviation σ).



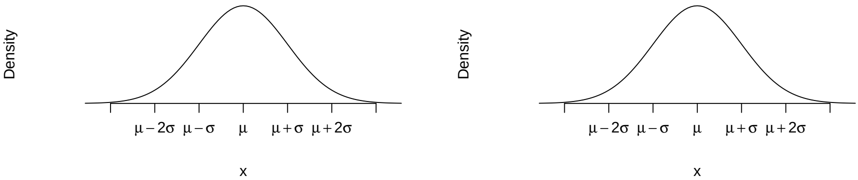
Properties of Normal Distribution

- The range of possible values of Y is $-\infty$ to ∞.
- The total area under the curve is 1.
- $\mu = E(Y)$ controls the center.
- The curve is symmetric around μ.
- The area under the curve above μ is 0.5 and the area below μ is 0.5.



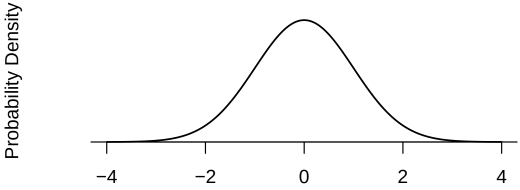
Properties of Normal Distribution

- $\sigma = \sqrt{Var(Y)}$ controls the spread.
- The area under the curve between $\mu - \sigma$ and $\mu + \sigma$ is about 2/3.
- The area between $\mu - 2\sigma$ and $\mu + 2\sigma$ is about 0.95.



The Standard Normal Distribution

- A random variable Z has a **standard normal distribution** if Z follows a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ (or, variance $\sigma^2 = 1$).
- The distribution curve is symmetric around 0.
- The area below/above 0 is 0.5.
- The area at an exact value z is 0.
- Linear transformation of normal random variables (r.v.'s) are still normal distributed.



Standardization to $N(0,1)$

- Let the random variable Y follow a general normal distribution with mean μ and variance σ^2.
- That is, $Y \sim N(\mu, \sigma^2)$.
- A useful transformation is

$$Z = \frac{Y - \mu}{\sigma} \sim N(0,1).$$

- For a possible value y, the term

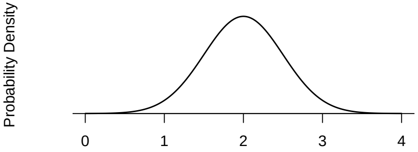
$$z = \frac{y - \mu}{\sigma}$$

is called the **z-score corresponding to** y.

Secchi Depth 1980 $P(Y < 0.25)$

- Let the random variable Y be the Secchi depth of a lake in 1980 randomly selected from $N(2.0, 0.5^2)$.
- What is the probability Y is less than 0.25 m?
- Since $Y \sim N(2.0, 0.5^2)$, apply the standardization

$$\begin{aligned} P(Y < 0.25) &= P\left(\frac{Y - 2.0}{0.5} < \frac{0.25 - 2.0}{0.5}\right) \\ &= P(Z < -3.5) \\ &= 0.0002 \end{aligned}$$



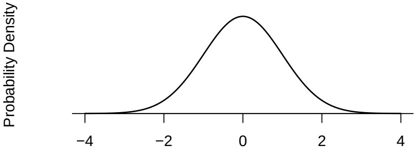
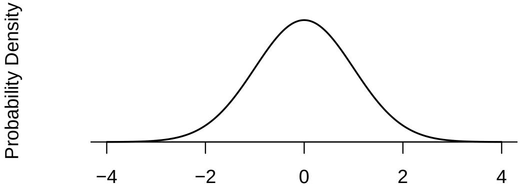


Table A for $P(Z > 1.5)$

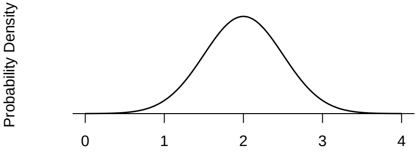
- Table A gives the upper tail probability $P(Z > z)$ for any given value z such that z is non-negative.
- The z values are on the outside and the probabilities are in the inside of Table A.
- Read the z value off x.x from the row index and 0.0x from the column index.
- Important: Drawing pictures helps!

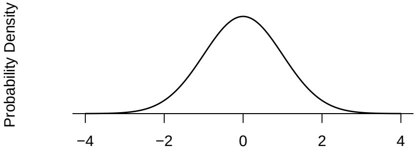


Secchi Depth 1980 0.90th Quantile

- Let the random variable Y be the Secchi depth of a lake in 1980 drawn from $N(2.0, 0.5^2)$.
- What value y would give probability $P(Y < y) = 0.90$?
- Find the 0.90th quantile of Z: $z = 1.282$.
- Note that

$$\frac{y - 2.0}{0.5} = 1.282.$$



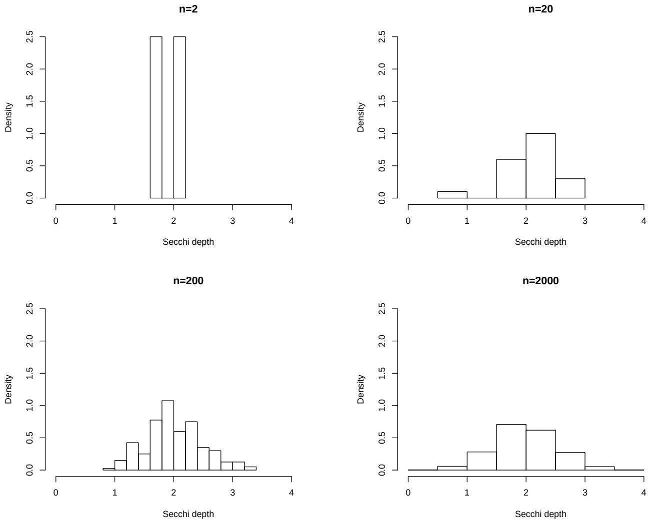


Outline

Example: Lake Clarity

- The population is comprised of 10,000 lakes and their Secchi depths in 1980.
- Previously we focused on drawing one lake at random from the population.
- We let random variable Y denote the Secchi depth value for a lake sampled at random from the population.
- Now, we consider drawing n lakes at random from the population.
- We let random variables $Y_1, Y_2, \ldots, Y_{n-1}, Y_n$ denote the Secchi depth values of n lakes sampled at random from the 1980 population.
- What is the effect of the sample size n?

Effect of Sample Size



Effect of Sample Size

n	mean	variance	SD
2	1.89	0.082	0.286
20	2.15	0.208	0.456
200	2.01	0.229	0.479
2,000	1.99	0.277	0.526
population	2.00	0.25	0.5

Adding or Subtracting Two Random Variables

- Considering adding two random variables $X + Y$ or subtracting one from the other $X - Y$.
- Example: $X =$ # of students in a randomly selected MS program and $Y =$ # of students in a randomly selected PhD program.
  - What is $E(X + Y)$?
- Example: $X =$ Secchi depth in 1990 and $Y =$ Secchi depth in 1980 of a randomly selected lake.
  - What is $E(X - Y)$?
- Also consider the spread of the new random variable $X + Y$ or $X - Y$.
  - What is $Var(X + Y)$? How about $Var(X - Y)$? What assumption do we need?

Expectation of $X + Y$ and $X - Y$

- The expectation of a sum is the sum of the expectations:

$$E(X + Y) = E(X) + E(Y)$$

- Alternative notation:

$$\mu_{X+Y} = \mu_X + \mu_Y$$

- The expectation of a difference is the difference of the expectations:

$$E(X - Y) = E(X) - E(Y)$$

- Alternative notation:

$$\mu_{X-Y} = \mu_X - \mu_Y$$

Variance of $X + Y$ and $X - Y$

- If X and Y are independent, then the variance of a sum is the sum of the variances:

$$Var(X + Y) = Var(X) + Var(Y)$$

- Alternative notation:

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$$

- If X and Y are independent random variables, then the variance of a difference is the *sum* of the variances:

$$Var(X - Y) = Var(X) + Var(Y)$$

- Alternative notation:

$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$$

- General formula?

Independence vs Dependence

- Two random variables X and Y are said to be **independent** if probability statements about X do not change, when we know the value of Y.
- If X and Y are independent, then it is also true that probability statements about Y do not change, when we know the value of X.
-
- For two events A and B, a **conditional probability of** A **given** B is denoted as $P(A|B)$.
- Two events A and B are **independent** iff $P(A|B) = P(A)$.
- When two events A and B are independent, then
  - $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$
  - $Cov(X, Y) = 0$

Example: Color and Shape of 7 Cards

- Example: At some point in time, the 8th card got lost.

card id	1	2	3	4	5	6	7
color	b	g	r	y	b	g	r
shape	$\square$	$\square$	$\square$	$\square$	$\triangle$	$\triangle$	$\triangle$

- Randomly draw a card from the box. What is the probability that it is blue?

$$P(\text{blue}) = \rule{8cm}{0.4pt}$$

- Given that the card randomly drawn from the box is a square, what is the probability that it is blue?

$$P(\text{blue}|\square) = \rule{8cm}{0.4pt}$$

Sample Mean

- Let $Y_1, Y_2, \ldots, Y_n$ denote an i.i.d. (identically and independently distributed) sample from a population with mean μ and variance σ^2.
  - The probability distribution of Y_i has mean μ and variance σ^2 for each $i = 1, 2, \ldots, n$.
  - The random variables $Y_1, Y_2, \ldots, Y_n$ are assumed to be independent of each other.
- Lake clarity example: As the sample size n increases, the sample mean gets closer to the population mean.
- The sample mean is defined as

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

- Note: $\bar{Y}$ is also a random variable.

Properties of Sample Mean

- Now, consider $Y \sim N(\mu, \sigma^2)$.
- Let $Y_1, Y_2, \ldots, Y_n$ denote an i.i.d. sample from this population $N(\mu, \sigma^2)$.
- The distribution of the sample mean $\bar{Y}$ is also normal.
- The expectation of the sample mean is ____________
- The variance of the sample mean is ____________
- The standard deviation (SD) of the sample mean is

$$\text{SD}(\bar{Y}) = \underline{\hspace{6cm}}$$

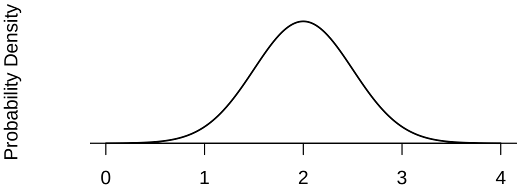
- Note: σ^2 is the population variance of Y.
- Alternative notation:

$$\mu_{\bar{Y}} = \mu_Y, \quad \sigma^2_{\bar{Y}} = \frac{\sigma^2_Y}{n}, \quad \sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{n}}.$$

- $\sigma_{\bar{Y}}$ is also known as the standard error of the sample mean.

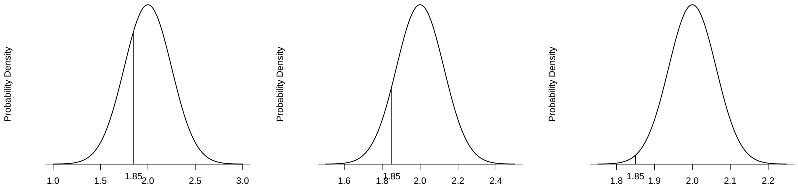
Computing Probabilities of Sample Mean

- Let $Y \sim N(\mu, \sigma^2)$ be the population.
- What is the distribution of $\bar{Y}$?
- What is the probability of $\bar{Y}$ between a and b? That is, $P(a < \bar{Y} < b) = ??$
- Example of Secchi depth 1980: $Y \sim N(2.0, 0.5^2)$.



Example: Secchi Depth 1980

- The population is well-approximated by $Y \sim N(2.0, 0.5^2)$.
- Take an i.i.d. sample of size $n = 4, 16, 64$.
- What is the probability of $\bar{Y}$ less than 1.85 m?



Central Limit Theorem

- Consider a random sample $Y_1, Y_2, \ldots, Y_n$, where Y_i's are independent and identically distributed (i.i.d.) random variables with mean μ and variance σ^2.
- Denote $Y_i \sim_{\text{i.i.d}} D(\mu, \sigma^2)$ for $i = 1, \ldots, N$. Note that the distribution D does not have to be Normal!
- Central limit theorem (CLT): f the sample size n is large enough, then the distribution of $\bar{Y}$ is closely approximated by a normal distribution with mean μ and variance σ^2/n. In other words, $\bar{Y}$ is approximately $N(\mu, \frac{\sigma^2}{n})$. Why $\frac{\sigma^2}{n}$?
- How large a sample is large?
  - If D is fairly symmetric and unimodal, then $n = 10$ or 20 may be enough.
  - If D is high skewed, then $n = 1000$ or more may be needed.

Discrete Random Variable

- So far, we have focused on random variables that follow normal distributions.
- A normal random variable Y is **continuous**, because the possible values of Y are from $-\infty$ to ∞.
- A random variable is **discrete** if there are a finite number of possible values or at most there is one for every integer.
- Toss a coin *independently* three times and record the number of times that the coin lands on heads.
- Let Y denote the number of heads.
- Then, Y is a discrete random variable.

Example: Coin Toss

- The possible outcomes are:

1st toss	2nd toss	3rd toss	Y
H	H	H	3
H	H	T	2
H	T	H	2
H	T	T	1
T	H	H	2
T	H	T	1
T	T	H	1
T	T	T	0

- Let π denote the probability of heads. Then $1 - \pi$ is the probability of tails.
- It can be shown that, under the independence assumption,

$$P(Y = 3) = \pi^3, \quad P(Y = 2) = 3\pi^2(1 - \pi),$$
$$P(Y = 1) = 3\pi(1 - \pi)^2, \quad P(Y = 0) = (1 - \pi)^3$$

Binomial Distribution

- A **binomial distribution** arises when the following conditions are satisfied:
  1. There are repeated trials, each of which can result in one of two outcomes, either “success” or “failure”;
  2. the probability of a success is constant for all trials, and is equal to π; the probability of a failure is $1 - \pi$;
  3. the trials are independent.
- Define a random variable Y to be the number of successes in n trials.
- Then Y is said to follow **a binomial distribution with parameters n and π** or Y is a **binomial random variable.**
- This is abbreviated as:

$$Y \sim B(n, \pi)$$

Binomial Distribution

- Binomial distribution is an important probability model and is often very useful.
- Examples:
  - Cure rate of a new drug
  - Proportion of lakes that are suitable habitats for a fish species
- However, the binomial distribution is not suitable for all situations.

Formula for $P(Y = y)$

- Let $Y \sim B(n, \pi)$. What is the probability that $Y = y$ for $y = 0, 1, \dots, n$?
- The formula for the binomial distribution is:

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

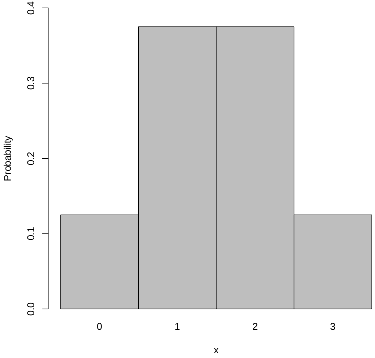
where

- $y = 0, 1, 2, \dots, n$ are the possible numbers of successes.
- $\pi^y = \underbrace{\pi \times \pi \times \cdots \times \pi}_{y \text{ times}}$.
- $(1 - \pi)^{n-y} = \underbrace{(1 - \pi) \times (1 - \pi) \times \cdots \times (1 - \pi)}_{n-y \text{ times}}$.
- n choose y:

$$\binom{n}{y} = \frac{n!}{y!(n - y)!}$$

Example: Binomial Distribution

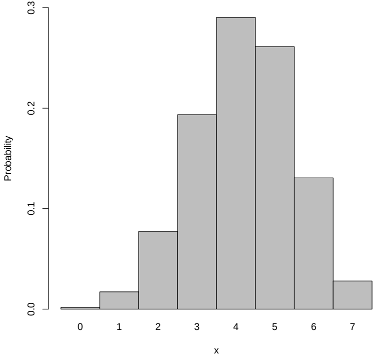
- Suppose $Y \sim B(3, 0.5)$.
- Probability histogram.



- What is $P(Y = 1)$?

Example: Binomial Distribution

- Suppose $Y \sim B(7, 0.6)$.
- Probability histogram.



- What is $P(Y = 3)$?

Properties of Binomial Distribution

- Let $Y \sim B(n, \pi)$.
- The expectation of Y is $\mu_Y = E(Y) = n\pi$
- The variance of Y is $\sigma_Y^2 = Var(Y) = n\pi(1 - \pi)$
- Suppose $Y \sim B(10, 0.25)$. What is the expectation and variance of Y?

- Suppose $Y \sim B(100, 0.25)$. What is the expectation and variance of Y?

Properties of Proportion of Successes

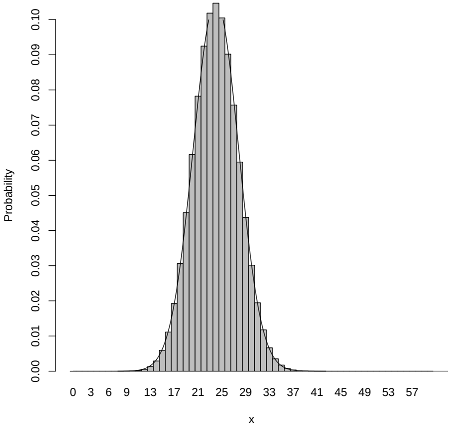
- In some situations, one is interested in the **proportion of successes** among n trials.
- Define $Y^* = \frac{Y}{n}$ where $Y \sim B(n, \pi)$.
- Then $E(Y^*) = \pi, \quad Var(Y^*) = \frac{\pi(1-\pi)}{n}$.

- Alternative notation: $\hat{\pi} = \frac{Y}{n}$
- Suppose $Y \sim B(10, 0.25)$. What is the expectation and variance of $\hat{\pi}$?

- Suppose $Y \sim B(100, 0.25)$. What is the expectation and variance of $\hat{\pi}$?

Example: $Y \sim B(60, 0.4)$

- Consider $Y \sim B(60, 0.4)$. Thus $n = 60, \pi = 0.4$.
- What is μ_Y and σ_Y^2?
- What is $P(Y \leq 20)$?



Normal Approximation of Proportion of Successes

- Consider proportion of success:

$$\hat{\pi} = \frac{Y}{n}$$

- For $Y \sim B(60, 0.4)$,

$$E(\hat{\pi}) = \pi = 0.4, \quad Var(\hat{\pi}) = \frac{\pi(1-\pi)}{n} = \frac{0.4 \times 0.6}{60} = 0.004.$$

- Hence by the CLT on an average,

$$\hat{\pi}_{\mathrm{NA}} \sim N(0.4, (0.0632)^2).$$

- What is $P(\hat{\pi} \leq 0.48)$?

Sample Variance

- Lake clarity example: As the sample size n increases, the sample variance gets closer to the population variance.
- Again, consider $Y \sim N(\mu, \sigma^2)$.
- Let $Y_1, Y_2, \ldots, Y_n$ denote an i.i.d. sample from this population $N(\mu, \sigma^2)$.
- The sample variance is defined as

$$S^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}$$

- Note that S^2 is also a random variable.

Properties of Sample Variance

- The expected value of the sample variance is

$$E\left(S^2\right) = \sigma^2$$

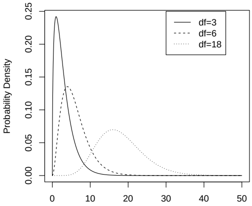
- Interpretation: ____________________________.
- The variance of the sample variance is

$$Var(S^2) = \frac{2\sigma^4}{n-1}.$$

- Interpretation: ____________________________.

Chi-Squared Distribution

- Next goal: Make probability statements about S^2.
- A **chi-squared distribution with** m **degrees of freedom** is a model for a random variable denote by $V^2 \sim \chi^2_m$.
- Properties of $V^2 \sim \chi^2_m$:
  - The range of possible values of V^2 is from 0 to $+\infty$.
  - The distribution is right-skewed.
  - $E(V^2) = m$ and $Var(V^2) = 2m$.



- If $Z_1, \ldots, Z_m$ are independent $N(0, 1)$, then $\sum_{j=1}^{m} Z_j^2 \sim \chi^2_m$.

Standardization of Sample Variance

- Suppose $Y_1, Y_2, \ldots, Y_n$ is an i.i.d. sample from a normal distribution with mean μ and σ^2.
- Let S^2 be the sample variance.
- Define
$$V^2 = \frac{(n-1)S^2}{\sigma^2}$$
- V^2 follows a chi-squared distribution with $n-1$ degrees of freedom.
- $\bar{Y}$ and S^2 are independent.

A Review

So far, we have

- used the probability distribution of a random variable Y to characterize the population.
- viewed the sample observations (i.e., data) as outcomes of random variables from this population.
- considered an i.i.d. sample $Y_1, Y_2, \ldots, Y_n$ of sample size n, where Y_i denotes the random variable for the ith observation in the sample.
- studied the distribution of the sample mean $\bar{Y}$ and the sample variance S^2.
- learned that $\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$ and $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$, assuming $Y_i \sim_{\text{i.i.d}} N(\mu, \sigma^2)$.

We will next turn to *statistical inference*.