Introductory Statistics Refresher

Dr. Julia L. Sharp

Short Course on Introductory Statistics
Part II

Beginning Probability

- Probability is used to help us make statistical inferences.
- As an example, suppose that I claim that I am excellent free throw shooter, making 80% or more of my free throw shots.

You watch me shoot 10 free throws and I only make 2 of them.

Do you agree that I'm free throw shooter?

The result of 2 out of 10 made free throws is highly improbable if my claim were true.

- Probability will help us to make judgements concerning the parameters of interest.

Useful Definitions

- Experiment: A process by which an observation is obtained (different from types of studies previously discussed)
- Sample Space (S): The set of all possible outcomes or results of an experiment.
- Simple event (E_i): An element of the sample space that cannot be decomposed into any smaller events.
- Complement ($\overline{E}_i$): The event that E_i does not occur.

Definition Example

Suppose we have a jar that contains a penny (P), two nickels (N_1,N_2), a dime (D), and a quarter (Q). Three coins are chosen without replacement. The order in which the coins are chosen is not important.

- Sample Space (S)

$S = \{(P, N_1, N_2), (P, N_1, D), (P, N_2, D), (N_1, N_2, D), (P, N_1, Q), (P, N_2, Q), (N_1, N_2, Q), (N_1, D, Q), (N_2, D, Q), (P, D, Q)\}$

0.11, 0.16, 0.16, 0.20, 0.31, 0.31, 0.35, 0.4, 0.4, 0.36

- Simple event (E_i)

$(P, D, Q) \leftarrow P(E_i) = 1/10$

- Complement ($\overline{E}_i$)

$(P, N_1, N_2), (P, N_1, D), (P, N_2, D), (N_1, N_2, D), (P, N_1, Q), (P, N_2, Q), (N_1, N_2, Q), (N_1, D, Q), (N_2, D, Q)$

$P(\overline{E}_i) = 1 - 1/10 = 9/10$

Probability Definitions

- Probability: A measure of belief that an event will occur on the next repetition of an experiment.
- Classical Interpretation of Probability: All outcomes in a sample space are equally likely.
  - $P(E) = \dfrac{\text{the total number of ways } E \text{ can occur}}{\text{the total number of equally likely events}}$ — # of E_i in S
  - Suppose that we roll two fair die. The probability of obtaining a sum of 7 is: $P(\text{sum of } 7) = \dfrac{6}{36} = \dfrac{1}{6}$ (1,1) (1,2)
- Relative Frequency Interpretation of Probability (Empirical Approach): The experiment is repeated a large number of times to estimate the probability.
  - $P(E) = \dfrac{\text{the number of times } E \text{ occured}}{\text{the maximum number of times } E \text{ could occur}}$
  - Suppose we simulate flipping a coin to determine the probability of obtaining a head. We flip the coin 1000 times and obtain a head 495 times: $P(\text{head}) = \dfrac{495}{1000}$. This seems to be a fair coin.

Probability Laws

Suppose E_i is an event in sample space S.

$0 \leq P(E_i) \leq 1$

- The probability for each E_i in S is between 0 and 1.
- The sum of the probabilities of all E_i in S is 1.
- The probability of the complement of simple event E_i is one minus the probability of the simple event. $P(\overline{E}_i) = 1 - P(E_i)$
- Two events E_1 and E_2 are mutually exclusive (disjoint) if when the experiment is performed a single time, then the occurrence of one of the events excludes the possibility of the occurrence of the other event. $P(E_1 \text{ AND } E_2) = 0$
- Two events E_1 and E_2 are independent if the occurrence of event E_1 is not dependent on the occurrence of event E_2 (and vice versa). $P(E_1 \text{ AND } E_2) = P(E_1) * P(E_2)$

Example of Probability Laws

E_1 is mutually exclusive of E_3, but they are not indep.

- Let E_1 be the event that we choose a penny. Let E_2 be the event that we choose 2 nickels. Let E_3 be the event that we choose 40 cents.

$P(E_1) = P(\text{choose a penny}) = 6/10 = 0.6$

$P(E_2) = 3/10 = 0.3$

$P(E_3) = 2/10 = 0.2$

Are E_1 and E_2 mutually exclusive?

$P(\text{penny and 2 nickels}) = 1/10 = 0.1$

E_1 and E_2 are not mutually exclusive

Are E_1 and E_2 independent?

$P(E_1 \text{ and } E_2) = 0.1 \stackrel{?}{\neq} P(E_1) \cdot P(E_2)$

$= 0.6 \cdot 0.3 = 0.18$

E_1 and E_2 are not independent

Variables: Discrete and Continuous

- Random variables: A numerical valued function defined on a sample space.

Y, X

- Discrete random variables: A random variable that can assume a countable number of outcomes.

Y = the amount of money taken from the jar if 3 coins are selected (countable number of outcomes)

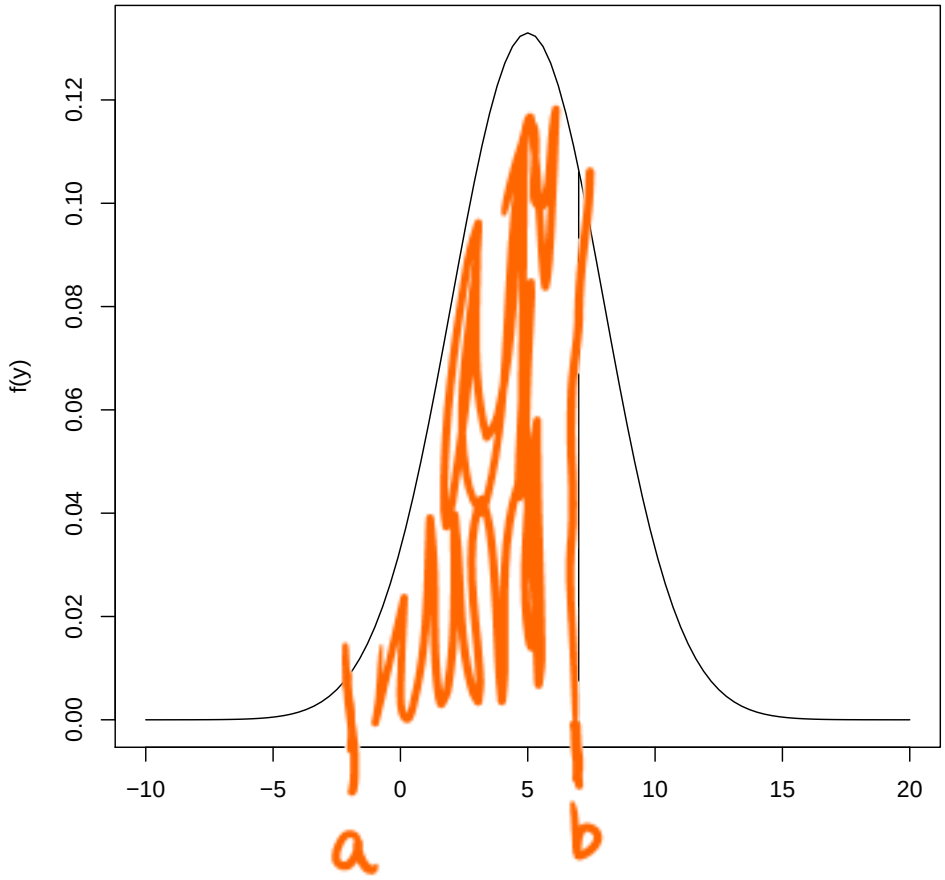
- Continuous random variables: A random variable that will assume an infinitely large number of values corresponding to the points on a line interval.

X = heights of individuals

Continuous Random Variable Properties

- Consider a relative frequency histogram with small class intervals.
- A smooth curve provides a model for the population relative frequency distribution.
- Histogram relative frequencies are proportional to areas and probabilities over the class intervals.
- Relative frequencies sum to one. The total area under the curve is one.

Continuous Random Variable Interval Probabilities



$P[a < Y < b]$

- $f(y)$ is the height of a curve for a given value of y.
- The probability that a continuous random variable falls in an interval between two numbers a and b is equal to the area under the curve over the interval.

$P[Y = y] \neq f(y)]$

The Normal Distribution

- Suppose Y is a normally distributed random variable. The height of the probability distribution for a specific value of y is represented by $f(y)$:

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

Smooth curve that is bell-shaped and symmetric around the mean μ

Properties of Normal Distributions

- Area: The area under the curve is one.
- Symmetry: The distribution is symmetric around the mean.
- Center: The mean and median are equal.
- Probabilities for events which come from a normal distribution may be found by determining the appropriate area under the curve.

There are many normal distributions. Even normal distributions with the same mean could have different variance.

The Standard Normal Distribution

- If x is an observation from a normal distribution with mean μ and standard deviation σ, the standardized value of x is

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

- z-score:
  - Tells us how many standard deviations the original observation is from the mean.
  - Observations more than the mean are positive when standardized.
  - Observations less than the mean are negative when standardized.
  - If a variable X follows a normal distribution ($N(\mu, \sigma^2)$), then the standardized variable $Z = \frac{X - \mu}{\sigma}$ follows a $N(0, 1)$ distribution.

The Standard Normal Table

prob. are inside the table

$P[Z \le z]$ ← observed

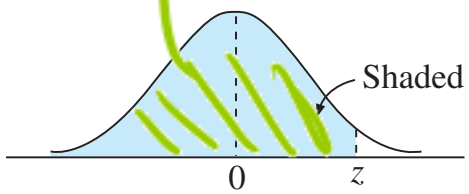


TABLE 1

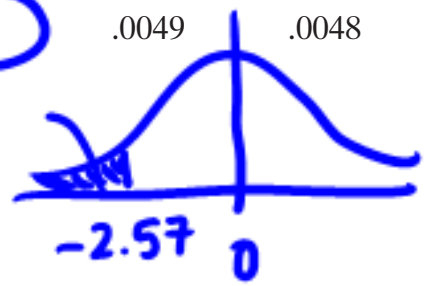
Standard normal curve areas

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

$P[Z < -2.57] = 0.0051$

$P[Z < a] = 0.0034$

$a = -2.71$



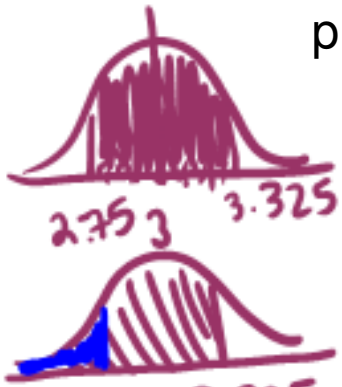
General Normal Distribution Word Problem

- A machine that cuts corks for wine bottles operates in such a way that the distribution of the diameter of the corks produced is normally distributed with mean of 3 *cm* and a standard deviation of 0.1 *cm*. Y = diameter of corks
  - Find the probability that a randomly selected cork will have a diameter smaller than 2.4 *cm*.

$Y \sim N(3, 0.1^2)$

$$P(Y < 2.4) = P\left(Z < \frac{2.4 - 3}{0.1}\right) = P(Z < -6) \approx 0$$

  - The specifications call for corks with diameters between 2.75 and 3.325 *cm*. A cork not meeting the specifications is considered defective. What proportion of corks produced by this machine are defective?



$$P(2.75 < Y < 3.325) = P(Y < 3.325) - P(Y < 2.75)$$

$$= P\left(Z < \frac{3.325 - 3}{0.1}\right) - P\left(Z < \frac{2.75 - 3}{0.1}\right)$$

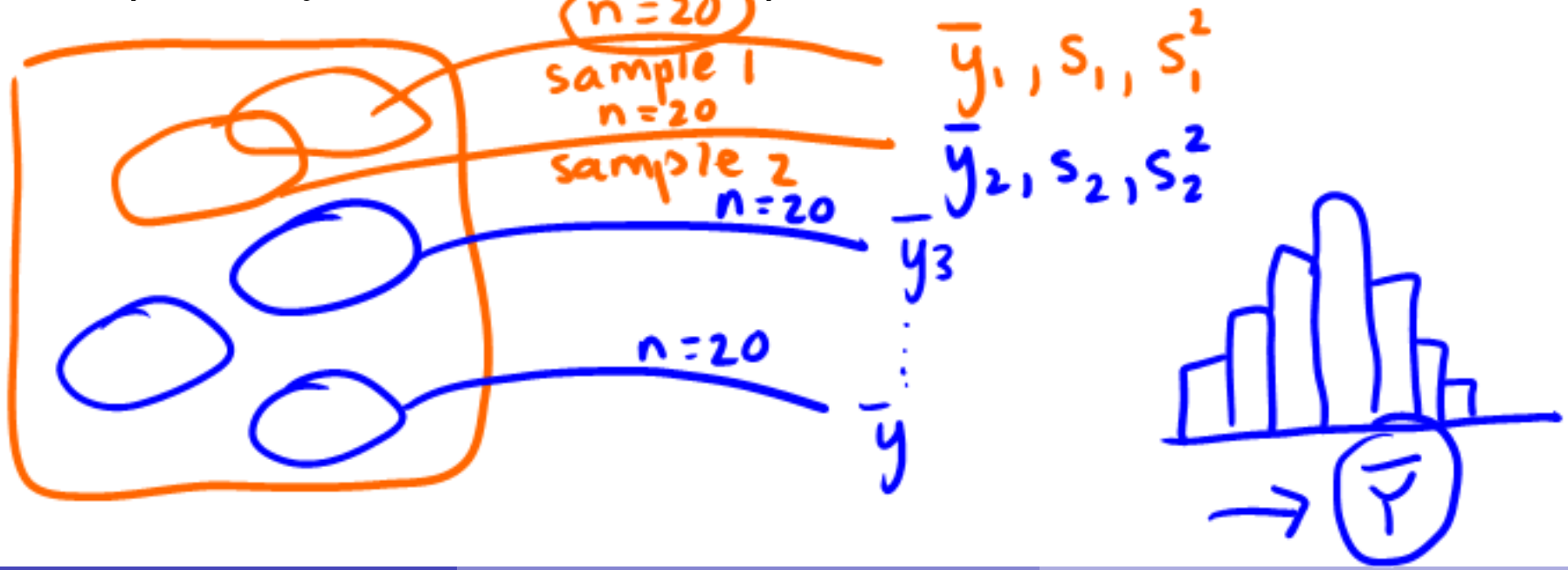
$$= P(Z < 3.25) - P(Z < -2.5) = 0.9994 - 0.0062 = 0.9932$$

1 − 0.9932 = 0.0068 are defective

= 0.9932 not defective

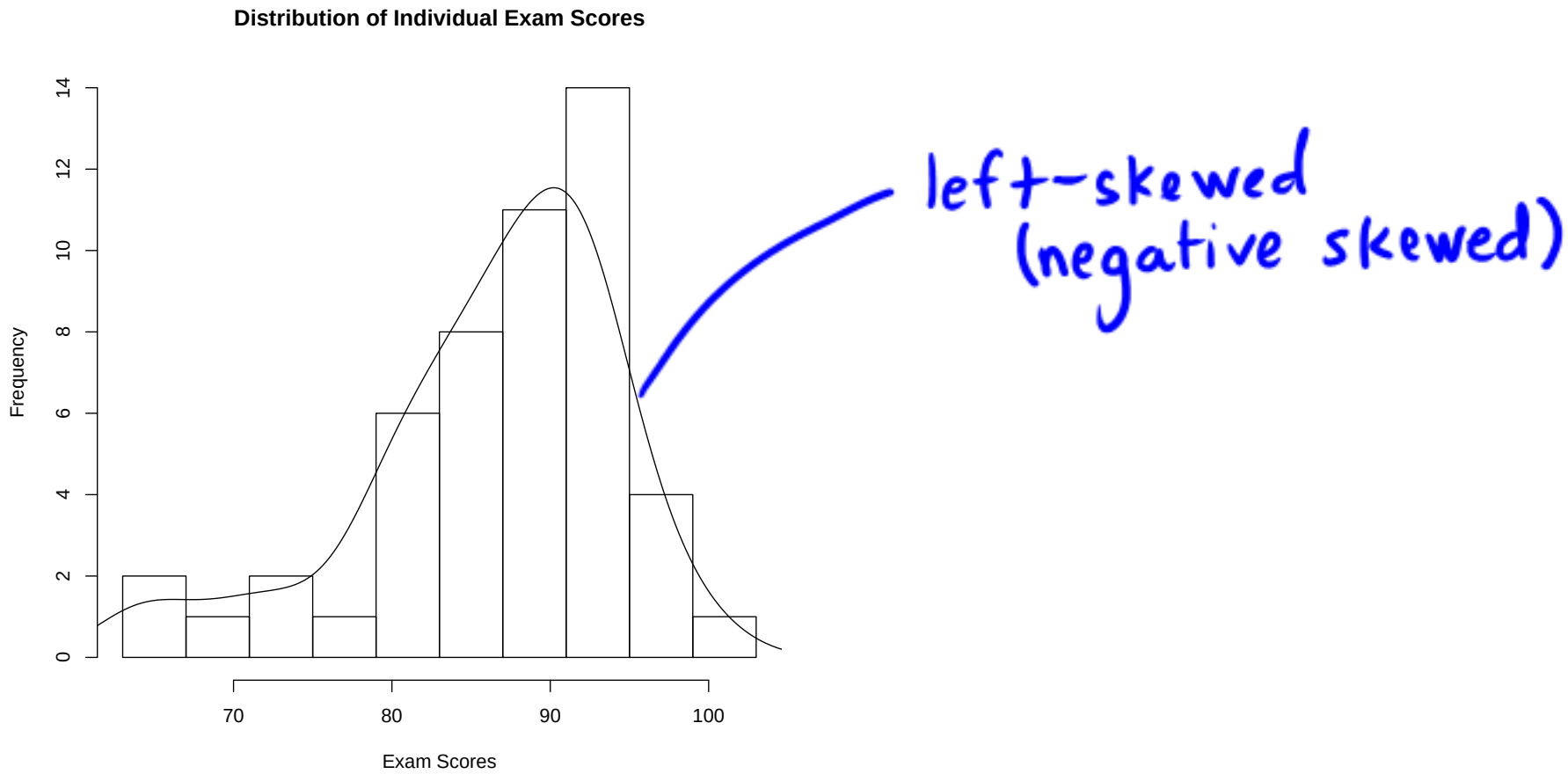
Sampling Distribution Definition

- Population parameters are fixed, unknown quantities.
- Statistics vary depending upon the sample selected (i.e., statistics are random variables).
- Sampling Distribution: the probability distribution for the values of the statistic that results when random samples of size n are repeatedly drawn from the population.



Sampling Distribution Example

- Suppose we have 50 exam scores for students in a statistics course. The random variable $Y =$ exam scores.
- The mean of the scores is $\mu = 86.16$.
- The standard deviation of the scores is $\sigma = 8.049236$.
- The distribution of exam scores is shown below:

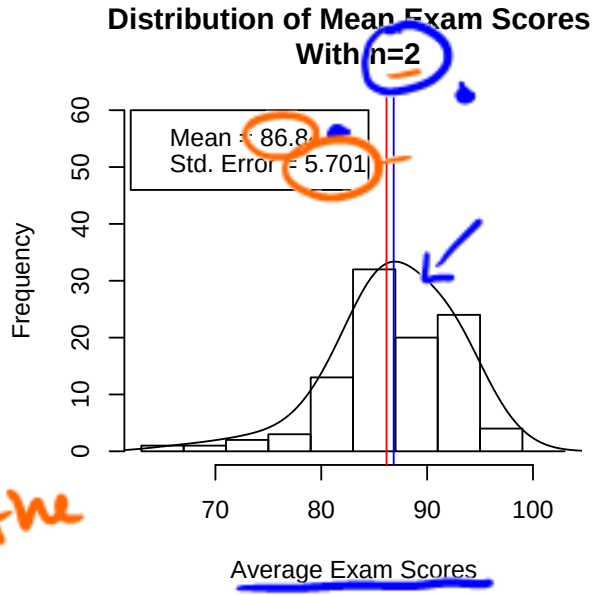


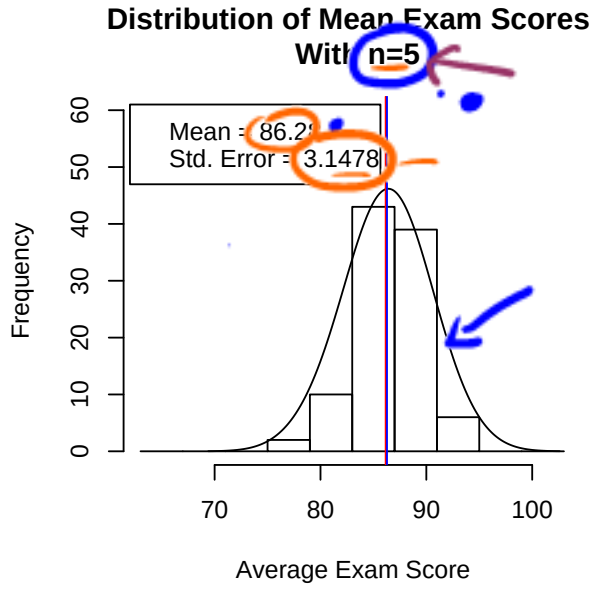
Sampling Distribution for Varying *n*

$\bar{Y}$ = average exam score

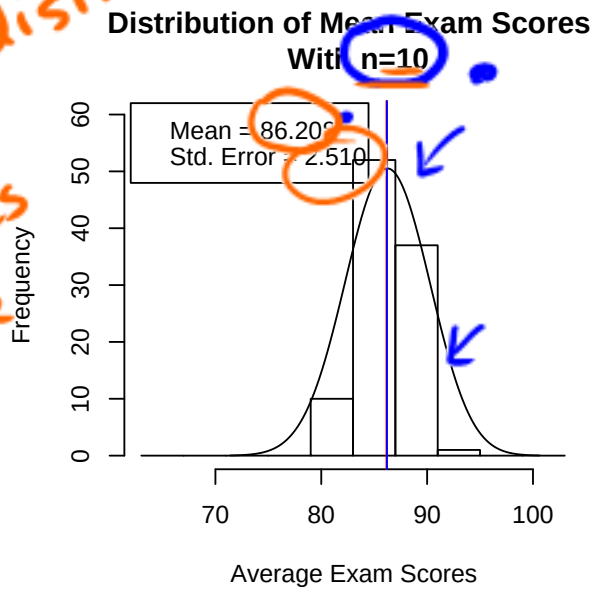
$\mu = 86.16$

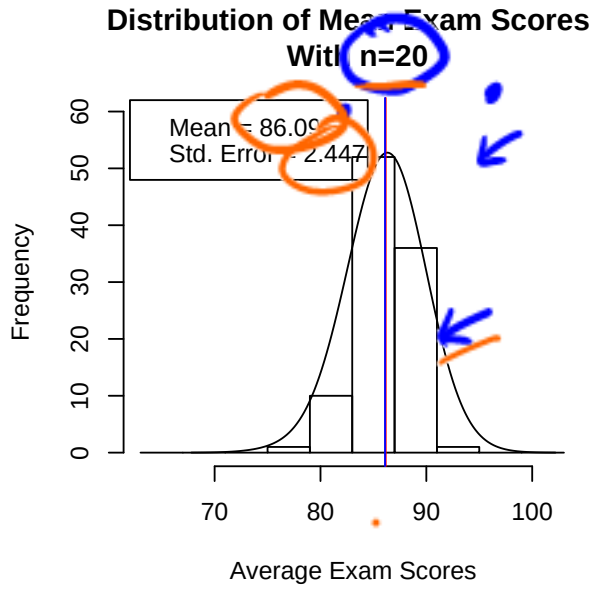
$\sigma = 8.05$





mean of samp. dist. about the same as the mean of the pop. dist.





std. dev. decreases as sample size increases

sampling dist. becomes approx. symm. with increasing n

Standard Error

- If all possible samples of size n from a population with a mean of μ and standard deviation σ are selected, the sampling distribution of the mean has mean μ and standard deviation $\sigma/\sqrt{n}$.
- Standard error: the standard deviation of the sampling distribution of the statistic
- The Central Limit Theorem: If random samples of size n are selected from a population with a finite mean, μ, and finite variance, σ^2, then when n is large, the distribution of the sample mean will be approximately normal with mean μ and variance σ^2/n. The approximation becomes more and more accurate as n becomes large.

$Y \sim (\mu, \sigma^2)$

If n is large, $\bar{Y} \dot{\sim} N(\mu, \sigma^2/n)$

Sampling Distribution Example

$Y \sim N(3, 0.1^2)$

- A machine that cuts corks for wine bottles operates in such a way that the distribution of the diameter of the corks produced is normally distributed with mean of 3 *cm* and a standard deviation of 0.1 *cm*.

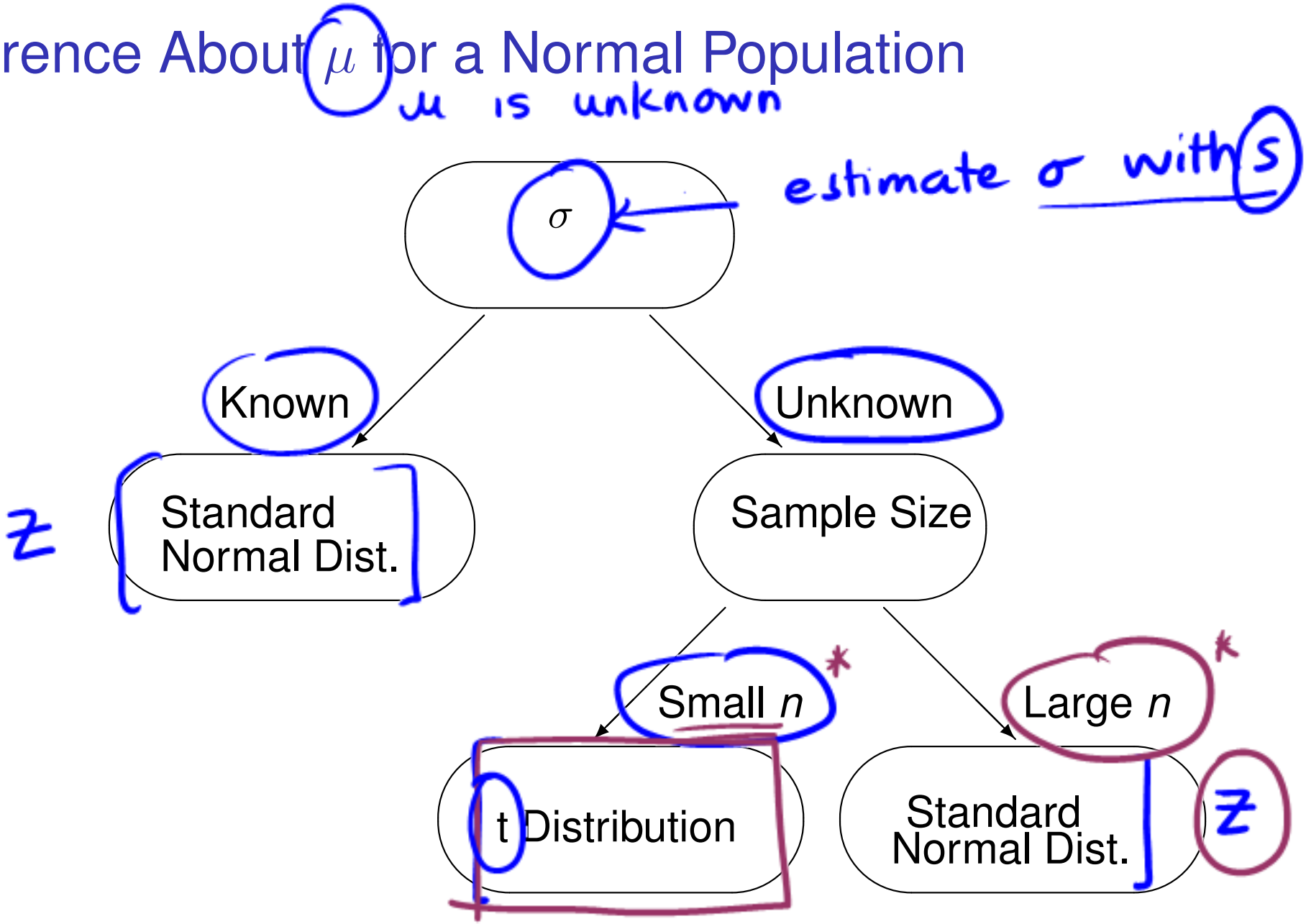
$\bar{Y} \sim N(3, 0.1^2/10)$

  - Suppose that a random selection of 10 corks for wine bottles is selected. Find the probability that the average diameter of corks is less than 2.9 *cm*.

$$P(\bar{Y} < 2.9) = P\left(Z < \frac{2.9 - 3}{0.1/\sqrt{10}}\right)$$

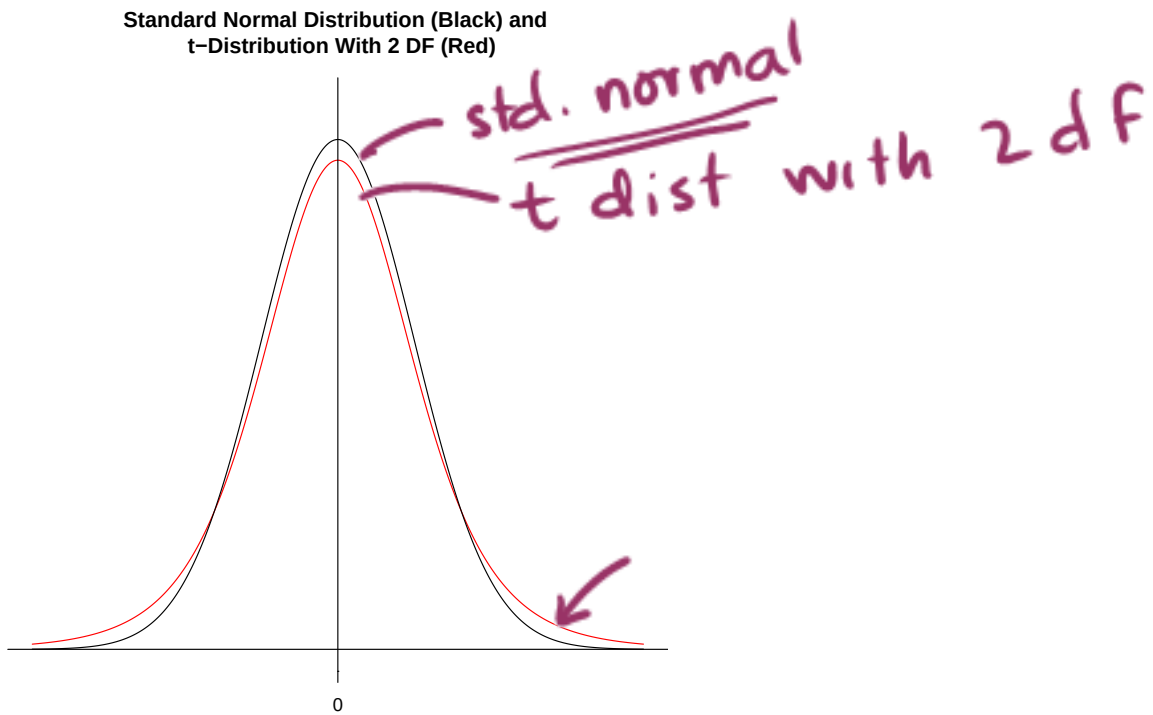
$$= P(Z < -3.16) = 0.0008$$

Inference About μ for a Normal Population



Properties of a *t* Distribution

- The mean is zero.
- The distribution is symmetric about the mean.
- The distribution has variance greater than 1; the variance approaches one as the sample size n becomes large.
- The distribution is less peaked at the mean and thicker at the tails than the normal distribution.
- The distribution approaches the standard normal distribution as the sample size becomes large.



t-Distribution Table

$$t = \frac{\bar{y} - \mu}{s/\sqrt{n}} \quad *$$

$$s = \sqrt{\frac{\sum y_i^2 - (\sum y_i)^2/n}{n-1}}$$

degrees of freedom

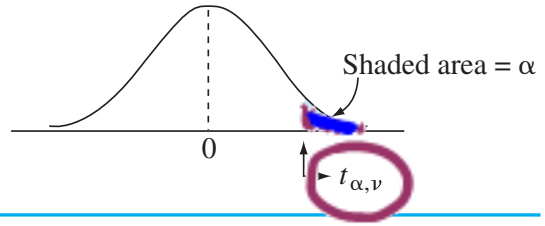


TABLE 2

Percentage points of Student's t distribution

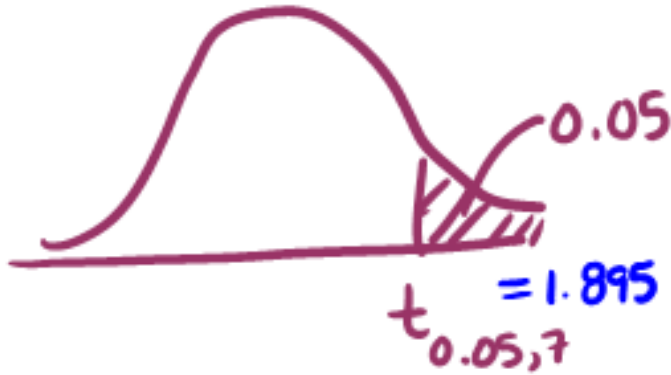
	Right-Tail Probability (α)								
α / ν	.40	.25	.10	.05	.025	.01	.005	.001	.0005
1	.325	1.000	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	.289	.816	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	.277	.765	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	.271	.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	.267	.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	.265	.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	.262	.706	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	.261	.703	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	.260	.700	1.372	1.812	2.228	2.764	3.169	4.144	4.587

prob.

t-scores (critical values)

Using the t Distribution Table

- $t_{0.05,7} =$



- ~~$t_{0.1,20} =$~~

0.005 < $P(T_{15} > 2.75)$ < 0.01

- $P(T_4 < -9)$ < 0.0005

Statistical Inference: an informed guess about a population parameter

- Estimation – estimate the value of a population parameter
- Testing – test a hypothesis about the value of a parameter

Estimation of μ

- Estimator: a rule that tells us how to calculate an estimate based on sample information
- Point Estimator: a rule that estimates the parameter with a single value
- Unbiased Estimator: an estimator whose average value is equal to the parameter being estimated

$\mu_{\bar{Y}} = \mu_Y$

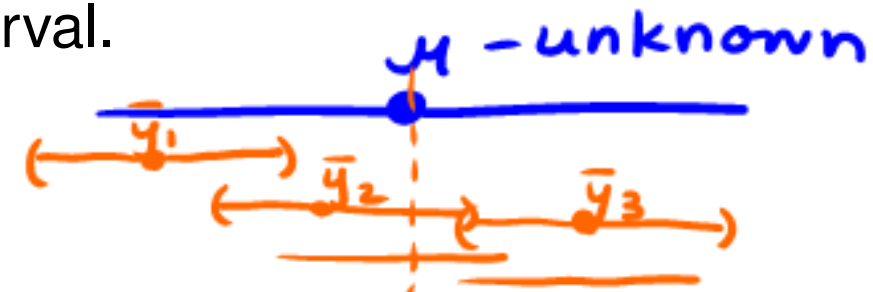
$\bar{Y}$ is unbiased for μ

- Estimate: a number calculated using an estimator

$\bar{Y}$ is a point estimator of μ

Confidence Intervals

- Interval estimator: a rule that provides two numbers that form an interval which is likely to contain the parameter. The interval is called a confidence interval.

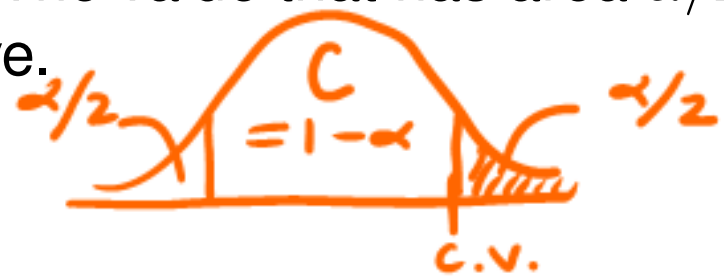


- Confidence coefficient: the fraction of times, in repeated sampling, the interval estimates encompass the parameter to be estimated.

95% of the intervals would contain μ

$C = 1 - \alpha$ 90% confident $C = 0.9$ $(\alpha = 0.1)$

- Critical value: The value that has area $\alpha/2$ to the right of it under the curve.



$(1-\alpha)100\%$ Confidence Interval for μ

- General interval:

estimate $\pm$ critical value * std. error

margin of error

- Interval for μ: $\bar{y} \pm$ c.v. * std. error

If σ is unknown, n is small

c.v. $= t_{\alpha/2, n-1}$ std. error $= s/\sqrt{n}$

If σ is unknown, n is large

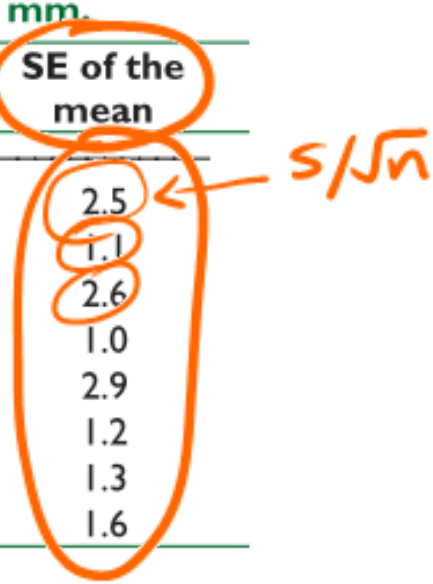
c.v. $= z_{\alpha/2}$ std. error $\sigma/\sqrt{n}$

Example

Riddle and Bergström (2013) describe several experiments to examine Phosphorus leaching from two soils. A table of results from one of the experiments is reproduced below. There were four different rain simulations used and two soil types (clay and sand). The amount of drainage water collected from lysimeters was recorded.

Table 4. The mean amount of drainage water (±SE, n = 1–5) collected from the lysimeters from each of the two soils and four rainfall simulations. The desired application was 50 mm.

Rainfall simulation	Soil	Mean	SE of the mean
		mm	
1	clay	37.6	2.5
	sand	27.1	1.1
2	clay	42.1	2.6
	sand	46.4	1.0
3	clay	47.3	2.9
	sand	46.2	1.2
4	clay	49.5	1.3
	sand	52.0	1.6



Riddle, M. U. and Bergström, L. (2013). "Phosphorus leaching from two soils with catch crops exposed to freeze-thaw cycles," *Agronomy Journal*, 105(3): 803-811.

Confidence Interval for Average Phosphorus Leaching

95% CI $\quad \bar{y} \pm c.v. * SE \quad n = 32$

```
        One Sample t-test

data:  drain$drainage
t = 32.4017, df = 31, p-value < 2.2e-16
alternative hypothesis: true mean is
                        not equal to 0
95 percent confidence interval:
 41.90197 47.53131
sample estimates:
mean of x
 44.71664
```

The TTEST Procedure

Variable: drainage

N	Mean	Std Dev	Std Err	Minimum	Maximum
32	44.7166	7.8069	1.3801	27.3436	56.7959

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
44.7166	41.9020	47.5313	7.8069	6.2588	10.3791

DF	t Value	Pr > |t|
31	-3.83	0.0006

$\bar{y} = 44.7166 \qquad s = 7.8069$

$$44.7166 \pm \underbrace{\left(t_{0.025,31}\right)}_{2.042} 7.8069/\sqrt{32}$$

$$(41.8984, 47.5348)$$

We are 95% confident that the average phosphorus leaching is between 41.9 and 47.5 mm.

Behavior of Confidence Intervals

- We want to balance the precision of the interval with the level of confidence.

- What happens to margin of error?

$$E = \text{c.v.} \cdot s/\sqrt{n}$$

margin of error decreases:
- when n is large
- when c.v. is smaller (less conf.)
- If s is small.

We can use this information to find what n should be for a particular margin of error