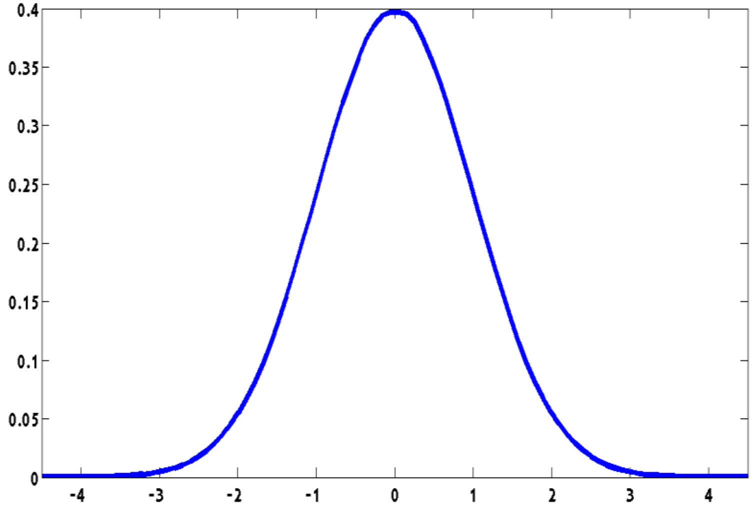
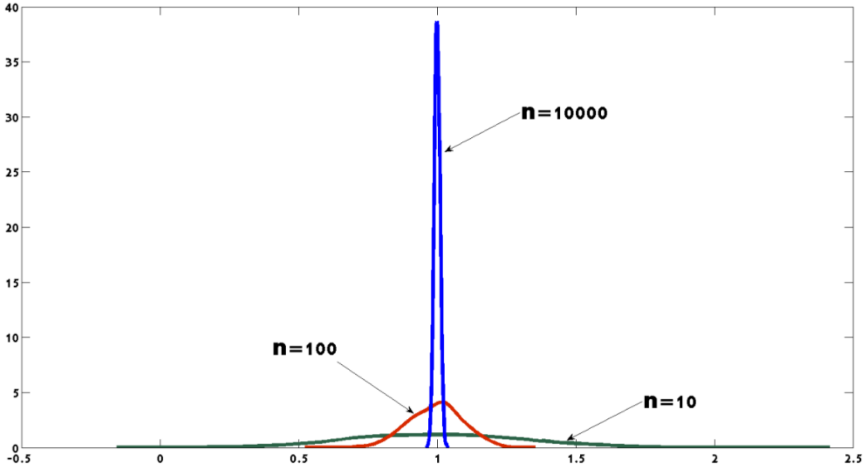
EC3303: Econometrics I

Review of Probability & Statistics
(Supplementary Lecture)





Kelvin Seah

AY 2022/2023, Semester 1

Random Variable (RV)

- A **random variable (RV)** is a variable whose value is an outcome of a random phenomenon.

- Random variables can be **discrete** or **continuous**.
  - discrete RV takes on only a discrete set of values like 0,1,2,…
  - continuous RV takes on a continuum of possible values.
  - E.g. number of times your computer crashes while you are writing a term paper is an example of a discrete RV.
  - Eg. length of time you take to write an email is an example of a continuous RV.

Probability Distribution

- Consider a discrete RV, X. Probability distribution of X lists the values and the probability that each value will occur:

Value of X	x_1	x_2	x_3	...	x_k
Probability	p_1	p_2	p_3	...	p_k

- p_i must satisfy 2 requirements:
  - Every p_i is a number between 0 and 1
  - $p_1 + p_2 + p_3 + \cdots + p_k = 1$

Distributions of Random Variables

- The **mean** is a measure of the **centre** of a distribution.

- The **variance/standard deviation** are measures of the **spread** or **dispersion** of a distribution.

Mean (Expected Value) of a RV

- The mean (expected value) of a random variable X – denoted $E(X)$ or μ_X – is an average of the possible values of X, but with a modification: modification to account for the fact that not all outcomes are equally likely.

- Suppose X is a discrete RV

Value of X	x_1	x_2	x_3	...	x_k
Probability	p_1	p_2	p_3	...	p_k

$$\mu_X = E(X) = x_1 p_1 + x_2 p_2 + x_3 p_3 + \cdots + x_k p_k = \sum_{i=1}^{k} x_i\, p_i$$

Standard Deviation & Variance of a RV

- **variance / standard deviation** measure the spread of a probability distribution.
- variance of a RV, X , is denoted $\sigma^2{}_X$ or $var(X)$

$$var(X) = E[(X - \mu_X)^2]$$

Value of X	x_1	x_2	x_3	…	x_k
Probability	p_1	p_2	p_3	…	p_k

$$\sigma^2{}_X = Var(X) = E[(X - \mu_X)^2] = (x_1 - \mu_X)^2 p_1 + (x_2 - \mu_X)^2 p_2 + \cdots + (x_k - \mu_X)^2 p_k$$

$$= \sum_{i=1}^{k} (x_i - \mu_X)^2 p_i$$

$$Var(X) = E[(X - \mu_X)^2]$$

- unit of the variance is awkwardly the unit of the square of X.
- standard deviation – denoted σ_X – is the square root of the variance. It is easier to interpret because it has the same units as X.

Rules for Means

- Let X , Y be RVs, a , b be constants, then:
- $E(a)=a$
- $E(a+bX)=a+bE(X)$
- $E(X+Y)=E(X)+E(Y)$ (expectation is a linear operator)

Rules for Variances

- Let $\boldsymbol{X}$, $\boldsymbol{Y}$ be RV's, $\boldsymbol{a}$, $\boldsymbol{b}$ be constants, then:
- ***var(a)=0***
- ***var(aX+b)=a²var(X)*** (note the squared constant)
- In general:

$$\mathbf{var}(X+Y)=var(X)+var(Y)+2cov(X,Y)$$
$$\mathbf{var}(X-Y)=var(X)+var(Y)-2cov(X,Y)$$

$$var(aX+bY)=a^2var(x)+b^2var(Y)+2abcov(X,Y)$$
$$var(aX-bY)=a^2var(x)+b^2var(Y)-2abcov(X,Y)$$

- but if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent,

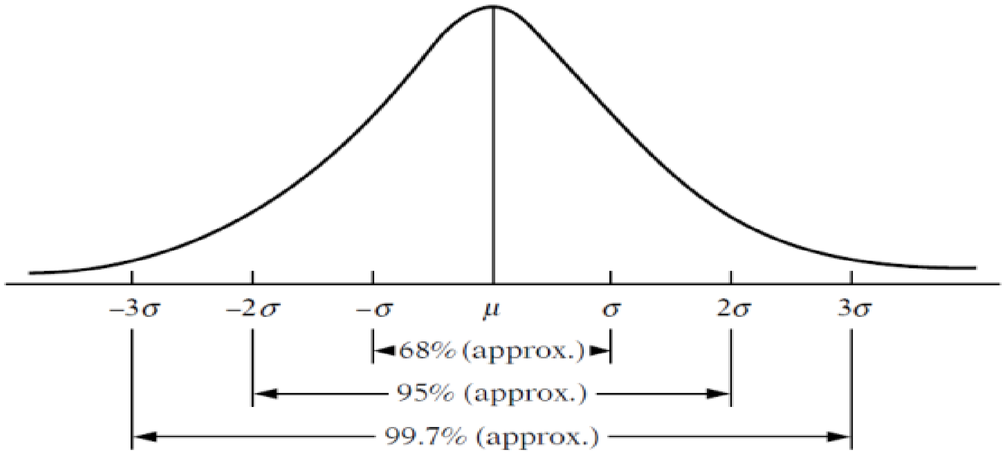
$$\mathbf{var}(X+Y)=var(X)+var(Y)$$

$$\mathbf{var}(X-Y)=var(X)+var(Y)$$
$$var(aX+bY)=a^2var(x)+b^2var(Y)$$
$$var(aX-bY)=a^2var(x)+b^2var(Y)$$

Normal Distribution

- Represents a large family of distributions, each with a unique mean & variance value.
- Suppose X follows a normal distribution with mean μ and variance σ^2

$$\boldsymbol{X \sim N(\mu, \sigma^2)}$$



- Approx 68% of the observations fall within σ of the mean μ .
- Approx 95% of the observations fall within 2σ of the mean μ .
- Approx 99.7% of the observations fall within 3σ of the mean μ .

- The normal distribution with mean = 0 and variance $\sigma^2 = 1$ is the **standard normal distribution**.

- RVs that have a standard normal distribution are denoted Z

$$\boldsymbol{Z \sim N(0, 1)}$$

- To compute probabilities involving normally distributed RVs,

  1. Standardize the variable.

  2. Look up standard normal c.d.f (Appendix Table 1).

Normal Distribution

- Suppose $X \sim N(1,4)$
  - How do we standardize X?
    - $Z = \frac{X-\mu}{\sigma} = \frac{X-1}{2} = \frac{1}{2}(x-1)$
    - $\frac{1}{2}(x-1) \sim N(0,1)$
    - can now use the Appendix Table 1 to answer questions like "what is the probability that $X \leq 2$?"

- Suppose $X \sim N(1,4)$
  - "What is the probability that $X \leq 2$?"
    - $\Pr(X \leq 2) = \Pr\left(\frac{X-1}{2} \leq \frac{2-1}{2}\right) = \Pr\left(Z \leq \frac{1}{2}\right) = 0.691$
  - "What is the probability that $1 \leq X \leq 2$?"
    - $\Pr(1 \leq X \leq 2) = \Pr\left(\frac{1-1}{2} \leq \frac{X-1}{2} \leq \frac{2-1}{2}\right)$

$$= \Pr\left(0 \leq Z \leq \frac{1}{2}\right) = \Pr\left(Z \leq \frac{1}{2}\right) - \Pr(Z \leq 0)$$

$$= 0.691 - 0.500 = 0.191$$

Review of Statistics

- Statistics is the science of using data to learn about unknown population distributions of interest.

- What is the mean of the distribution of earnings in Singapore?

  1) Perform an exhaustive survey of all workers in Singapore and construct the population distribution of earnings.

  or

  2) Select a random sample from the population of workers in Singapore. Then use statistical methods to draw inferences (**statistical inference**) about the full population.

  → Method (2) is more practical.

3 Ingredients of Statistical Inference

1. **Estimation** – computing a "best guess" numerical value for an unknown characteristic of a population distribution, from a sample of data

2. **Hypothesis testing** – formulating a specific hypothesis about the population, then using sample evidence to decide whether it is true.

3. **Confidence intervals** – computing an interval for an unknown population characteristic, using a sample of data

Review 3 concepts in the context of inference about an unknown population mean.

Estimation of the Population Mean

- You want to estimate the mean earnings μ_Y of the population of workers in Singapore. How would you do this?

- A possible way:

  - Choose a random sample of n workers.

  - Use the sample average $\bar{Y} = \frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n)$ to estimate the unknown population mean μ_Y.

  - $\bar{Y}$ is an example of an ***estimator*** of the population mean μ_Y.

Estimator vs Estimate

Estimator

- An **estimator** is a procedure / formula used to obtain an estimate of the parameter of interest
- It is a function of the (randomly drawn) sample of data.
- It is a RV, because it depends on a randomly selected sample.

Estimate

- An **estimate** is a numerical value of the estimator when it is computed using data from a specific sample.
- An estimate is just a number and so is nonrandom.

Estimator vs Estimate

- sample average

$$\bar{Y} = \frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n)$$

is an **estimator** of the population mean μ_Y.

- Suppose we have drawn a random sample of 500 Singaporeans aged between 18 and 65 and collected data on their earnings.

- Then, we use the above formula to compute the average income and find it is SG$35,100. This number is an **estimate.**

Sample of Data Drawn Randomly from a Population: $Y_1,\ldots,Y_n$

- Under simple random sampling (SRS),
  - n objects selected at random from a population & each member of the population is equally likely to be included in the sample.
  - n observations are $(Y_{1,} Y_{2,\ldots\ldots,} Y_n)$ where Y_1 is the value of the first observation, Y_2 is the value of the second observation, and so on.
  - Prior to sample selection, the value of each Y_i , $i = 1, \ldots, n$ is random & can take on many possible values. So each Y_i is a random variable.
  - Once the objects are selected & the values of Y are observed, each Y_i becomes a number – no more random.

I.I.D. Observations in the Dataset

- Because individuals are selected at random, knowing the value of Y_1 provides no information about Y_2. Thus:
  - Y_1 and Y_2 are ***independently distributed*****.**
  - Y_1 and Y_2 come from the same distribution, and so, Y_1 and Y_2 are ***identically distributed*****.**
    - distribution of each Y_i , where $i = 1, \dots, n$, is the same as the population distribution of Y**.**
  - Under SRS, $Y_1, Y_2, \dots, Y_n$ are independently & identically distributed (***i.i.d.***)

What an SRS Scheme is Not

- You want to know the unemployment rate in Singapore.

  So you survey people sitting in parks at 10am on a Tuesday.

- You want to know the mean age of all OCBC customers

  So you survey OCBC customers at 10.30am on a weekday.

Sampling Distribution of the Sample Average

- sample average $\bar{Y}$ of n observations $Y_1, \ldots, Y_n$:

$$\bar{Y} = \frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

- Drawing a random sample means that the sample average is itself a random variable
  - Since $Y_1, \ldots, Y_n$ are random variables, their average is also a random variable.
  - Had a different sample been drawn, the observations & the sample average would have been different.
  - The value of $\bar{Y}$ varies from one randomly drawn sample to the next.
    - i.e. value of $\bar{Y}$ will vary in repeated sampling.
  - Since $\bar{Y}$ is a random variable, it has a probability distribution – known as the sampling distribution.

Value of $\bar{Y}$ varies from one randomly drawn sample to the next…

- I am interested in knowing the mean height of students in this class.
- I can draw a random sample of size 3 (n=3) and compute the average height:

Sample 1	
	Height (cm)
1	
2	
3	
Average Height	

Sample 2	
	Height (cm)
1	
2	
3	
Average Height	

- The value of the sample average varies from sample to sample.

Mean & Variance of the Distribution of the Sample Average $\bar{Y}$

- The exact (finite-sample) sampling distribution of $\bar{Y}$ is determined by the sample size n & the population distribution.
- If the population has mean μ_Y & variance ${\sigma^2}_Y$,

1.

$$\boldsymbol{E(\bar{Y}) = \mu_{\bar{Y}} =}$$

2. since the observations are independent, covariance between the Y_i's are 0.

$$\boldsymbol{var(\bar{Y}) = {\sigma^2}_{\bar{Y}} =}$$

- Standard deviation of $\bar{Y}$ is the square root of the variance:

$$\boldsymbol{std.dev(\bar{Y}) = \sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{n}}}$$

- these results hold whatever the distribution of Y_i (also the population distribution) is

- Since the mean of $\bar{Y}$ is the same as the mean of the population:
  - The sample average $\bar{Y}$ is an **unbiased** estimator of the unknown population mean μ_Y.
  - More on unbiasedness soon

- Since $var(\bar{Y}) = \frac{{\sigma^2}_Y}{n}$, the variability of the sampling distribution of $\bar{Y}$ decreases as the sample size n grows.

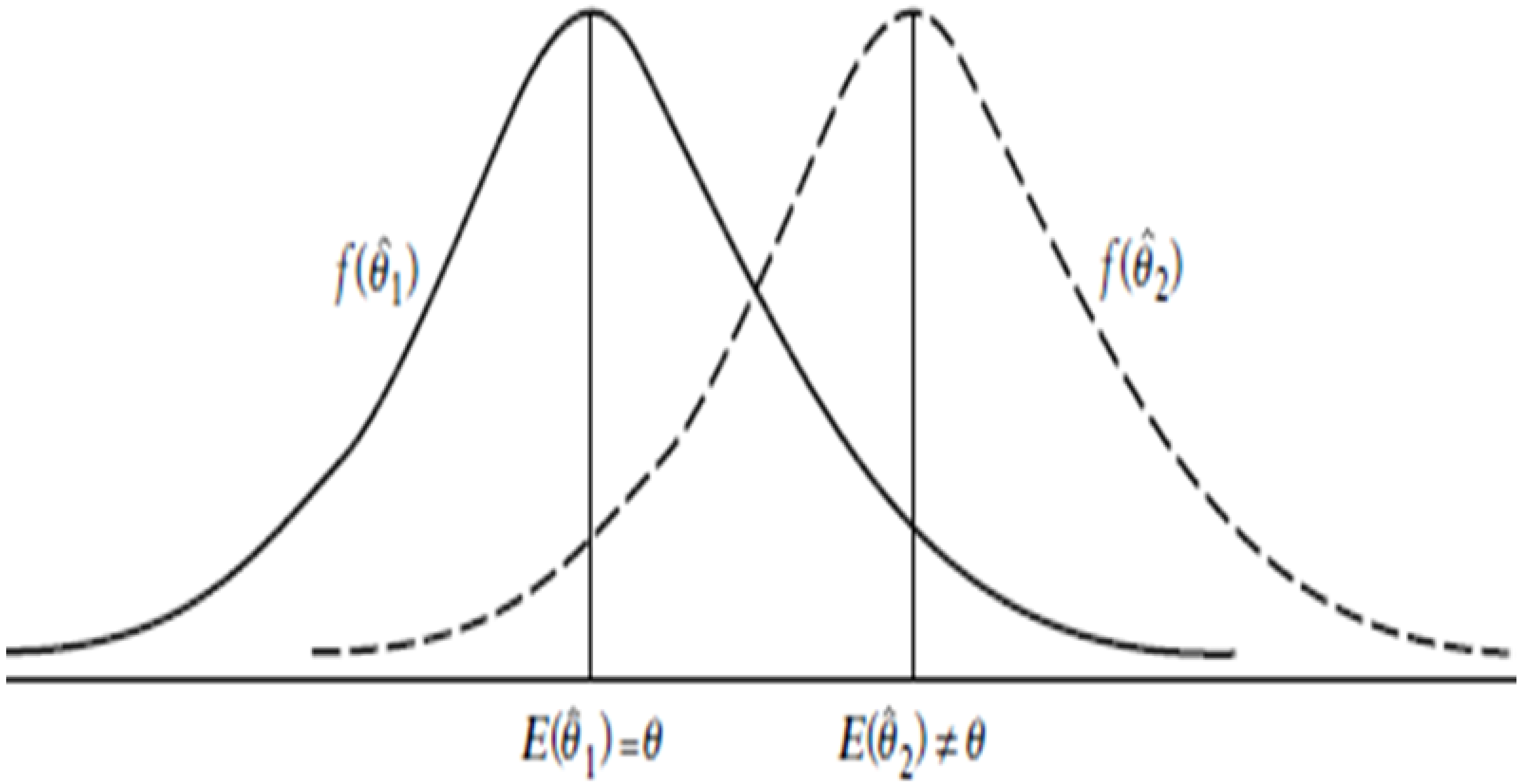
Unbiasedness

- Let $\hat{\theta}$ be an estimator of θ.
- We say that $\hat{\theta}$ is **unbiased** if

$$\boldsymbol{E\left(\hat{\theta}\right) = \theta}$$

  - if we draw samples repeatedly and take an average of the resulting estimates, we will get the true value – the estimator is correct **on average**.
- $\boldsymbol{E\left(\hat{\theta}\right) - \theta}$ is called **bias**.
- Since $E(\bar{Y}) = \mu_Y$, the sample mean is an unbiased estimator of the population mean.

Unbiased vs Biased Estimator



Shape of the Sampling Distribution of $\bar{Y}$

- Exact shape of the distribution of $\bar{Y}$ depends on the shape of the population distribution
  - if the population distribution is normal, then so is the distribution of $\bar{Y}$.
- Let a population be distributed $N(\mu_Y, \sigma^2{}_Y)$, then the sample average $\bar{Y}$ of n independent observations has the $N(\mu_Y, \frac{\sigma^2{}_Y}{n})$ distribution
  - many population distributions are not normal however.
  - so, in general, the finite sample distribution of $\bar{Y}$ can be complicated.

Large-Sample Approximations to Sampling Distributions

- As the sample size n increases, the distribution of $\bar{Y}$ gets closer to a normal distribution. This result is true no matter what shape the population distribution has as long as the population has a finite variance (i.e. $\sigma^2{}_Y < \infty$).

Central Limit Theorem

Draw an SRS of size n from any population with mean μ_Y and finite variance $\sigma^2{}_Y$. When n is large, the sampling distribution of $\bar{Y}$ is approximately normal.

$$\boldsymbol{\bar{Y} \text{ is approximately } N(\mu_Y, \frac{\sigma^2{}_Y}{n})}$$

- How large must n be?
  - Quality of the normal approximation depends on the population distribution.
  - $n \geq 100$ is typically sufficient for a wide variety of population distributions.

Central Limit Theorem

As $n \to \infty$,

$$\bar{Y} \xrightarrow{d} N(\mu_Y, \frac{\sigma^2{}_Y}{n})$$

- $\bar{Y}$ is said to have an ***asymptotic normal distribution*** if the distribution of $\bar{Y}$ approaches the normal as n grows large.
- When n is large, the distribution of the standardized sample average $\frac{\bar{Y}-\mu_Y}{\sigma_{\bar{Y}}}$ is well approximated by a $N(0{,}1)$ distribution

$$\frac{\bar{Y}-\mu_Y}{\sigma_{\bar{Y}}} \xrightarrow{d} N(0, 1)$$

- So the asymptotic normal distribution of $\frac{\bar{Y}-\mu_Y}{\sigma_{\bar{Y}}}$ does not depend on the distribution of Y (population distribution)!

Law of Large Numbers & Consistency

- **Law of large numbers (LLN)** states that when the sample size n increases, $\bar{Y}$ will be near the population mean μ_Y with increasing probability

As $n \to \infty$,

$$\bar{\boldsymbol{Y}} \xrightarrow{\boldsymbol{p}} \boldsymbol{\mu_Y}$$

- $\bar{Y}$ is a **consistent** estimator of μ_Y.

- LLN says that if we can afford to keep on measuring more people, then we will eventually estimate the mean earnings of Singaporean workers very accurately.

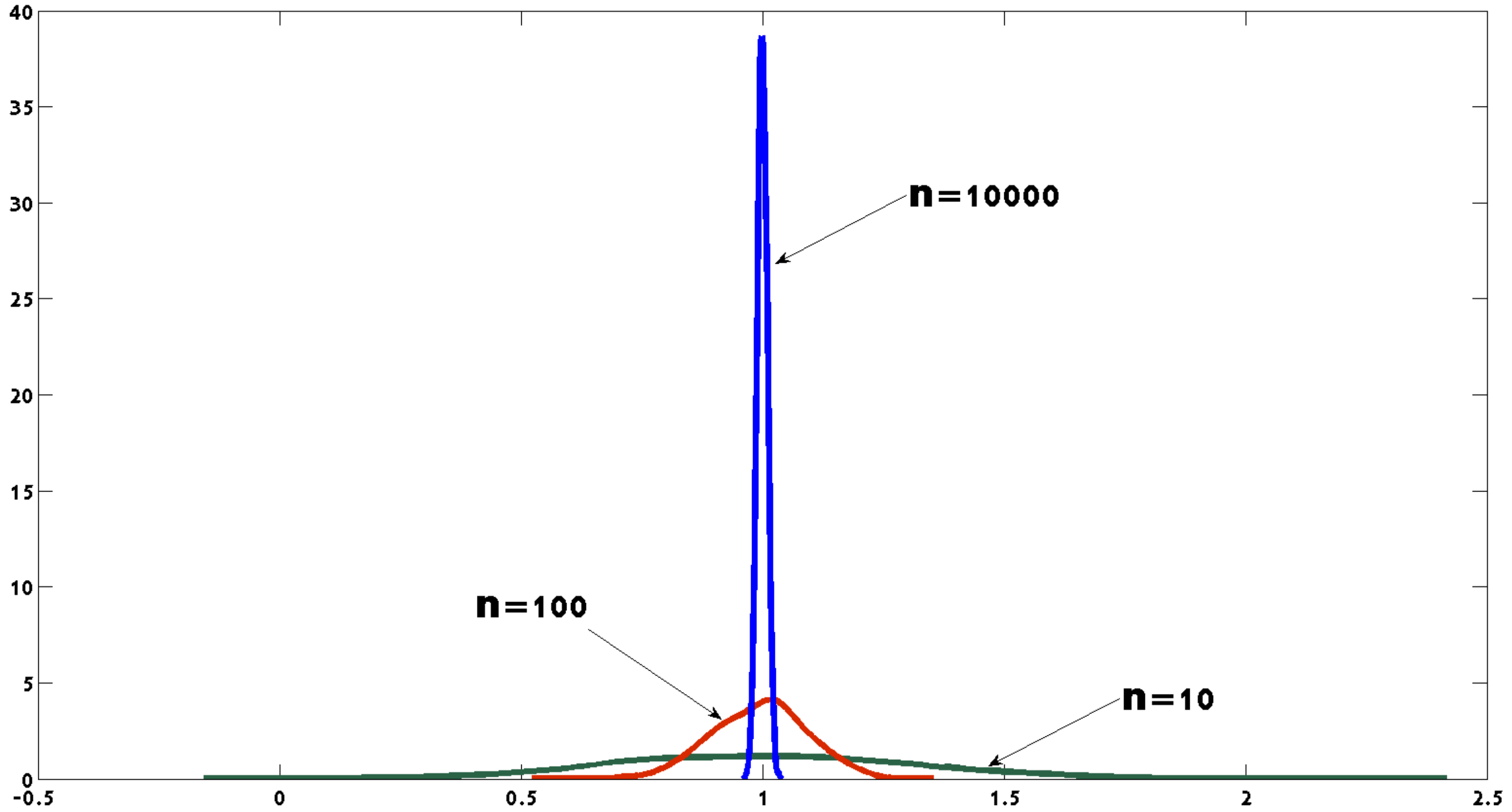
Consistency

- Let $\hat{\theta}$ be an estimator of θ.
- $\hat{\theta}$ is **consistent** if

$$\hat{\boldsymbol{\theta}} \xrightarrow{\boldsymbol{p}} \boldsymbol{\theta}$$

  - that is, if $\hat{\theta}$ is consistent, then the probability that it is within a small interval of the true value θ approaches 1 as the sample size increases.
  - For an estimator to be consistent, both bias & variance should tend to 0 as n gets large.
- $\bar{Y}$ is unbiased. Also, its variance, $var(\bar{Y}) = \frac{\sigma^2{}_Y}{n}$ approaches 0 as n gets large.
- So $\bar{Y}$ is a consistent estimator of the population mean μ_Y.

Consistency in a Picture



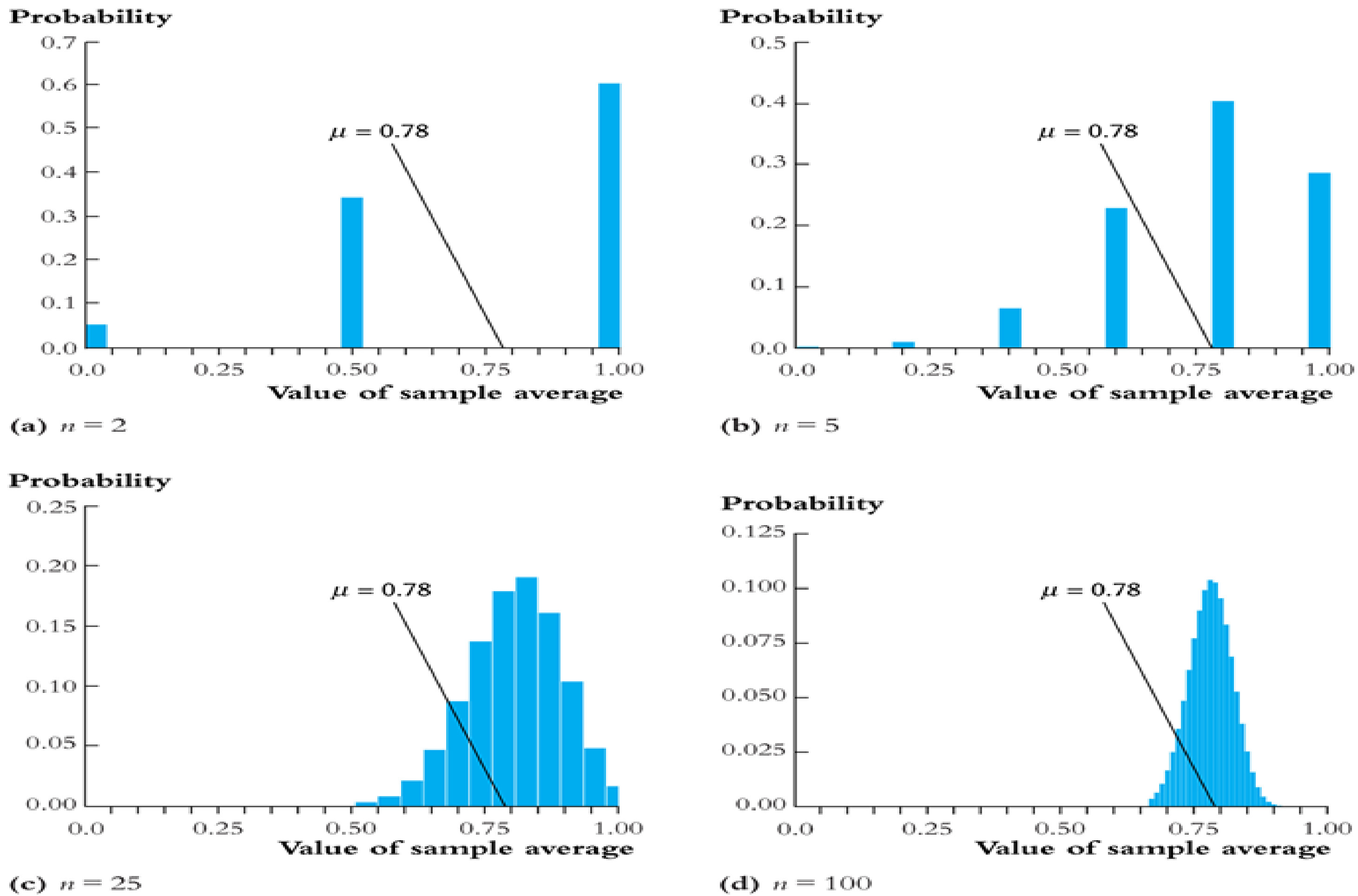
LLN & CLT at Work: An E.g.

- Let Y represent commute time, where $Y = 1$ if a commute is short & $Y = 0$ if it is long. Suppose Y is distributed:

Value of Y	0	1
Probability	0.22	0.78

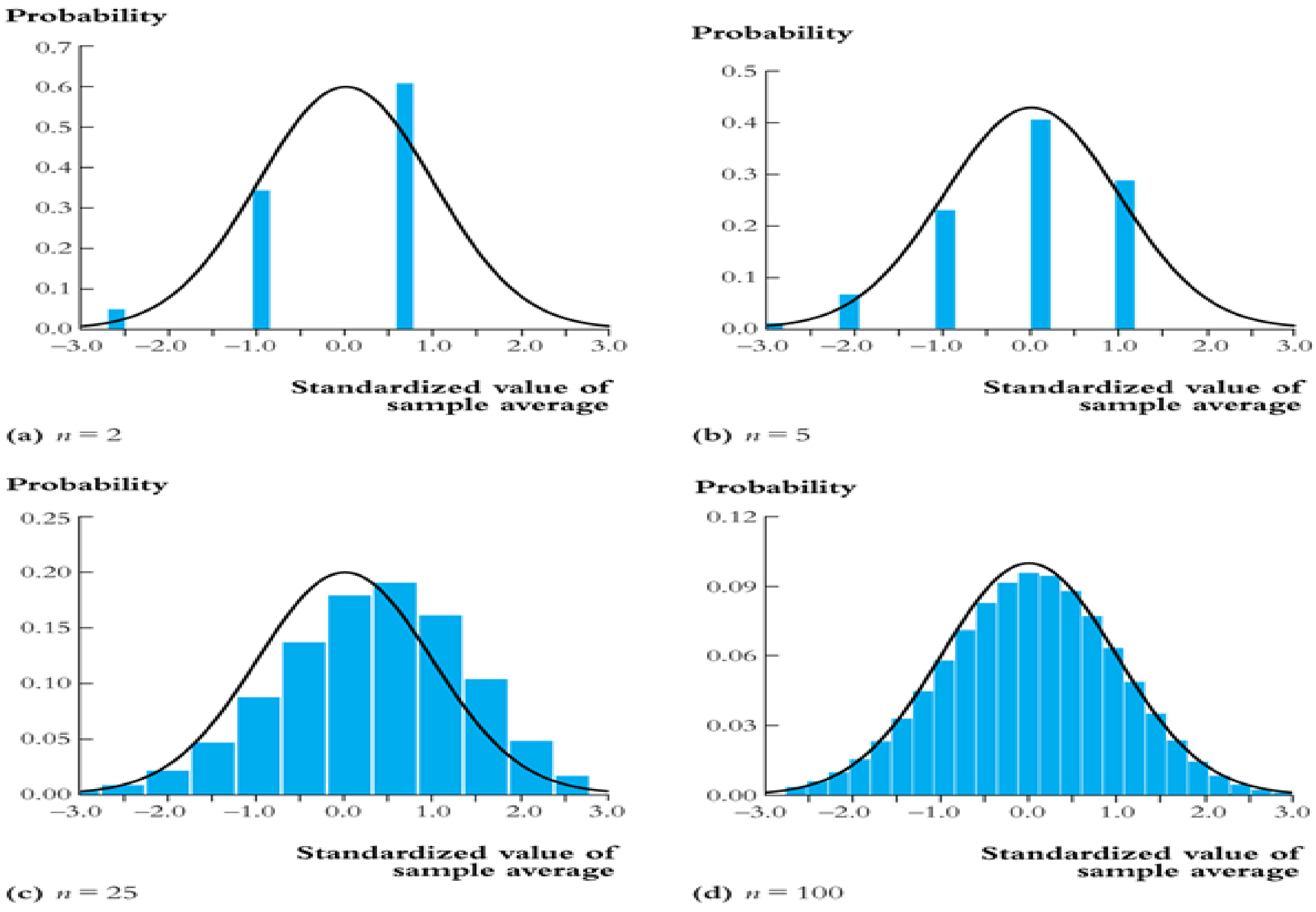
- Then $E(Y)$, in this e.g., tells us the probability of having a short commute on any randomly selected day.
  - $E(Y) = \mu_Y = E(Y_i) = 0.78$ is the fraction of commutes over a large number of commutes where the commute is short.
- In practice, the distribution of Y (and hence μ_Y) is unknown and has to be estimated.
  - can estimate μ_Y using $\bar{Y}$, where $\bar{Y}$ is the fraction of commutes in a sample in which the commute is short.

Law of Large Numbers and Consistency at Work



As n gets larger, the variance of the sampling distribution of $\bar{Y}$ decreases and the sampling distribution becomes more tightly concentrated around the true mean μ_Y=0.78

Central Limit Theorem at Work



As n gets larger, the sampling distribution of $\bar{Y}$ becomes increasingly well approximated by a normal distribution.

Summary of the Sampling Distribution of $\bar{Y}$

- Keep taking random samples of size n from a population with mean μ_Y and variance $\sigma^2{}_Y$.
  - Find the sample average $\bar{Y}$ for each sample
  - Collect all the $\bar{Y}$'s and display their distribution.
  - That's the sampling distribution of $\bar{Y}$.
- Sampling distribution of a sample average $\bar{Y}$ has mean μ_Y and variance $\frac{\sigma^2{}_Y}{n}$.
  - distribution of $\bar{Y}$ is normal if the population distribution is normal.
  - it is approximately normal for large n even if the population distribution is not normal.

Estimation of the Population Mean…Again

- $\bar{Y}$ provides one way to estimate μ_Y. But it is not the only way.
- For e.g., another way to estimate μ_Y is to use the first observation Y_1.
  - In repeated samples, Y_1 will take on different values, so Y_1 has a sampling distribution.
  - Because we assume SRS, the sampling distribution of Y_1 will be the same as the population distribution of Y.
  - So $E(Y_1) = \mu_Y$ and Y_1 is an unbiased estimator of μ_Y.
  - Since Y_1 and $\bar{Y}$ are both unbiased estimators of μ_Y, how do we choose between them?

Efficiency

- Let $\hat{\theta}$ and $\tilde{\theta}$ be unbiased estimators of θ. We prefer the estimator with the tighter sampling distribution. In other words, we prefer the estimator with the smaller variance.

- Suppose $\boldsymbol{var}(\widehat{\boldsymbol{\theta}}) < \boldsymbol{var}(\widetilde{\boldsymbol{\theta}})$,

  - then $\hat{\theta}$ is more **efficient** than $\tilde{\theta}$ and we prefer using $\hat{\theta}$ as our estimator.

First Observation Y_1 or Sample Average $\bar{Y}$?

- We prefer estimators which are more efficient.

$$var(Y_1) = \sigma^2{}_Y$$

$$var(\bar{Y}) = \frac{\sigma^2{}_Y}{n}$$

- For $n \geq 2$, $var(\bar{Y}) < var(Y_1)$, so $\bar{Y}$ should be used instead of Y_1.

- In fact, $\bar{Y}$ is actually the most efficient estimator among all unbiased estimators that are linear (i.e. linear estimators of μ_Y are weighted averages of $Y_{1,} Y_{2,\ldots\ldots,} Y_n$).

- $\bar{Y}$ is therefore **BLUE** (**Best Linear Unbiased Estimator**) of μ_Y.

Desirable Properties of Estimators

- We would like an estimator that gets as close as possible to the unknown true value (parameter).

- Hence the 3 desirable characteristics of an estimator are that it is:

  1) Unbiased

  2) Consistent

  3) Efficient

Hypothesis Tests

- We now know how to select an estimator with good properties.
- However, the result we get is just a point estimate. If we collected another sample & computed another estimate, the value of the estimate would likely be different. Can we say more about our result?
- Hypothesis testing allows us to say whether the value we got as an estimate is "compatible" with some hypothesized value about the population.
- E.g.: Given estimate of $31 per hour, can we say that the mean wage in Singapore is not $30 per hour?

Hypothesis Testing: Terminology

- **Null Hypothesis:** a hypothesis to be tested; usually a statement of "no effect" or "no difference". Denoted H_0.
- **Alternative Hypothesis:** a hypothesis we test the null against; this is the statement we hope is true if the null is not. Denoted H_1.
- A hypothesis is **simple** if it specifies a certain value for the parameter tested. For example,

$$H_0\colon\ \mu = 30$$

- Otherwise, it is a **composite.** For example,

$$H_1\!:\mu > 30 \qquad \text{or}$$
$$H_1\!:\mu \neq 30$$

- Null hypotheses are **always simple** ; Alternative hypotheses are **always composite.**

Hypothesis Tests Concerning the Population Mean

- Specify the null & alternative hypotheses, depending on your question

  $\boldsymbol{H_0: E(Y) = \mu_{Y,0}}$ **vs** $\boldsymbol{H_1: E(Y) \neq \mu_{Y,0}}$ **(2-sided alternative)**

  $\boldsymbol{H_0: E(Y) = \mu_{Y,0}}$ **vs** $\boldsymbol{H_1: E(Y) < \mu_{Y,0}}$ **(1-sided alternative)**

  $\boldsymbol{H_0: E(Y) = \mu_{Y,0}}$ **vs** $\boldsymbol{H_1: E(Y) > \mu_{Y,0}}$ **(1-sided alternative)**

  where $\mu_{Y,0}$ is the value of the population mean under the null hypothesis.

- Problem is to use the evidence in a randomly selected sample of data to decide whether to "accept" (not reject) the null hypothesis H_0 or to reject it in favour of the alternative H_1.

Test Statistics

- Hypothesis tests are based on a statistic which estimates the parameter of interest (E.g. estimate of $31 in wage example)

- If H_0 is true, we expect the estimate to take a value near the parameter value specified by H_0.

- Values of the estimate far from the value specified by H_0 give evidence against H_0. Alternative hypothesis determines which directions count as evidence against H_0.

<u>E.g.</u>

$H_1: E(Y) \neq 30$ – values of the estimate far from 30 give evidence against H_0.

$H_1: E(Y) < 30$ – only values of the estimate lower than 30 give evidence against H_0.

$H_1: E(Y) > 30$ – only values of the estimate greater than 30 give evidence against H_0.

- To assess how far the estimate is from the parameter specified by the null hypothesis, standardize the estimate:

$$test\ statistic = \frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}} = \frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}$$

- In practice, the population standard deviation σ_Y is unknown & we estimate it using the sample standard deviation s_Y
  - Using s_Y instead of σ_Y is possible because

$$s^2{}_Y \xrightarrow{p} \sigma^2{}_Y$$

where

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

- Using s_Y in place of σ_Y, we have the **t-statistic**

$$\boldsymbol{t = \frac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}}}$$

where $s_Y/\sqrt{n}$ is called the standard error of $\bar{Y}$ or $SE(\bar{Y})$.

- think of $t = \frac{\bar{Y}-\mu_{Y,0}}{s_Y/\sqrt{n}}$ as a standardized version of $\bar{Y}$ assuming the null hypothesis is true.
    - CLT says that when n is large, the t-statistic will have an approximate $N(0,1)$ distribution.
- **"Reject" or "do not reject" H_0 based on either the**

1) p-value *or*

2) pre-specified significance level.

Calculating the p-value (2-sided Alternative)

- p-value is the probability of drawing a statistic (estimate) which is as extreme or more extreme than the one computed using your sample of data, assuming the null hypothesis is true.

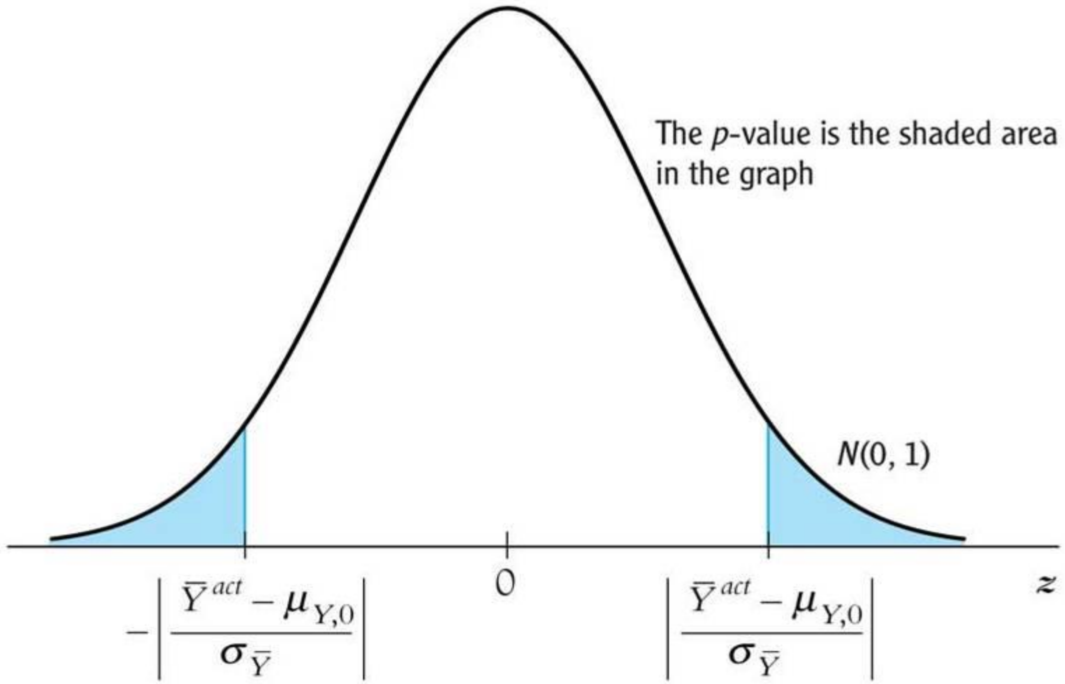
- Let $\bar{Y}^{act}$ denote the value of the sample average actually computed using the sample & Pr_{H_0} denote the probability computed assuming the null hypothesis is true, then:

$$p - value = Pr_{H_0}\left[\left|\bar{Y} - \mu_{Y,0}\right| > \left|\bar{Y}^{act} - \mu_{Y,0}\right|\right] = Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right|\right]$$

$$= Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{\frac{\sigma_Y}{\sqrt{n}}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\frac{\sigma_Y}{\sqrt{n}}}\right|\right] \cong Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{\frac{s_Y}{\sqrt{n}}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\frac{s_Y}{\sqrt{n}}}\right|\right]$$

$$\text{p-value} = 2\Phi(-|t^{act}|)$$

p-value: area in the tails of a standard normal outside $|t^{act}|$



- For large n, p-value = probability that the test-statistic falls outside $\left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right|$

- In practice, $\sigma_{\bar{Y}}$ is unknown – it must be estimated by $s_Y/\sqrt{n}$, the **standard error** of $\bar{Y}$.

Calculating the p-value: An e.g.

You want to test whether the average per hour earnings in Singapore are significantly different from \$30 at the 5% level. You randomly sample 100 residents and find:

$$\bar{Y}^{act} = \$33, \qquad s_Y = 9, \qquad n = 100$$

- Formulate hypothesis: $H_0: E(Y) = 30$; $H_1: E(Y) \neq 30$
- Compute the t-statistic: $t^{act} = \frac{\bar{Y}^{act} - \mu_{Y,0}}{s_Y/\sqrt{n}} = \frac{33-30}{9/\sqrt{100}} = 3\,\frac{1}{3}$
- Calculate the p-value: $2\Phi(-|t^{act}|) = 2\Phi\left(-\left|3\,\frac{1}{3}\right|\right) \approx 0.00$
- Since the p-value is so small (≈ 0.00), it is unlikely that such a sample would have been drawn if the null hypothesis is true. So conclude that null hypothesis is false.
- If the p-value computed was not small however (say 0.4), then it is quite likely that our observed sample average of \$33 could have arisen just by random sampling variation even if the null hypothesis were true. Accordingly, we cannot reject the null hypothesis.

Hypothesis Testing with a Prespecified Significance Level

- Reject H_0 when the p-value is small:
  - the smaller the p-value, the stronger the evidence against H_0 provided by the data.
- But how small is small?
  - must set a benchmark on how small the p-value should be before we reject H_0.
  - Fixed benchmark is called a **significance level**, α.

- When you conduct a hypothesis test, you can make 2 types of mistakes:

I. **Type I error:** rejecting the null when it is true.

II. **Type II error:** not rejecting a null when it is false.

- **Significance level, α:** prespecified probability of a type I error that we are willing to tolerate, e.g., 5% or 1%.
  - If you choose $\alpha = 0.05$, then you will reject the null if and only if p-value<0.05.
  - If you choose $\alpha = 0.01$, then you will reject the null if and only if p-value<0.01.

Hypothesis Tests Using a Fixed Significance Level

- Can perform hypothesis tests without computing p-values if we fix the significance level.

- Suppose we choose $\alpha = 0.05$, then for a 2-sided alternative, we reject the null if $|t^{act}| > 1.96$

  - 1.96 is called a **critical value**. It cuts off 5% of the area under the tails of the distribution of the t-statistic.

  - If we choose $\alpha = 0.01$, then for a 2-sided alternative, we reject the null if $|t^{act}| > 2.58$

  - If we choose $\alpha = 0.1$, then for a 2-sided alternative, we reject the null if $|t^{act}| > 1.64$

Summary of Hypothesis Testing with 2-sided Alternatives

- Compute the standard error of $\bar{Y}$, $SE(\bar{Y}) = s_Y/\sqrt{n}$

- Compute the t-statistic:

$$t^{act} = \frac{\bar{Y}^{act} - \mu_{Y,0}}{s_Y/\sqrt{n}}$$

- Compute the p-value:

$$\text{p-value} = 2\Phi(-|t^{act}|)$$

  - Reject null at the 5% significance level if the p-value<0.05.

  - Equivalently, reject null at the 5% significance level if $|t^{act}| > 1.96$

One-Sided Alternatives

- In some circumstances, we might have reason to think that the population mean exceeds the hypothesized value

$$H_1 : E(Y) > \mu_{Y,0} \quad \text{(1-sided alternative)}$$

- For a 1-sided alternative of this form, only large positive values of the sample average (and hence the t-statistic) count as evidence against H_0.

- So we reject H_0 only if the t-statistic takes on a large enough positive value.

$$H_1: E(Y) > \mu_{Y,0} \text{ (1-sided alternative)}$$

To test the one-sided alternative above:

- Compute the standard error of $\bar{Y}$, $SE(\bar{Y}) = s_Y/\sqrt{n}$ (as before)
- Compute the t-statistic: (as before)

$$t^{act} = \frac{\bar{Y}^{act} - \mu_{Y,0}}{s_Y/\sqrt{n}}$$

- Compute the p-value: (calculation of p-values is modified)

$$p - value = Pr_{H_0}(t > t^{act}) = 1 - \Phi(t^{act})$$

p-value here is the area under the standard normal distribution to the right of the calculated t-statistic.

- Reject H_0 at the 5% significance level if the p-value<0.05

$$H_1 : E(Y) > \mu_{Y,0} \text{ (1-sided alternative)}$$

- critical values for a 1-sided alternative are different:
  - Suppose we choose $\alpha = 0.05$, then for the 1-sided alternative above, we reject the null if $t^{act} > 1.64$.
  - It cuts off 5% of the area under the **upper tail** of the distribution of the t-statistic.
  - Suppose we choose $\alpha = 0.01$, then for the 1-sided alternative above, we reject the null if $t^{act} > 2.33$.
- If the alternative hypothesis is $H_1 : E(Y) < \mu_{Y,0}$, everything discussed applies except that signs are switched.

One-Sided Alternative: An e.g.

You want to test whether the average per hour earnings in Singapore are significantly more than \$30 at 5% level. You randomly sample 100 residents and find:

$$\bar{Y}^{act} = \$31, \qquad s_Y = 8, \qquad n = 100$$

- Formulate hypothesis: $H_0: E(Y) = 30$; $H_1: E(Y) > 30$
- Compute the t-statistic: $t^{act} = \frac{\bar{Y}^{act} - \mu_{Y,0}}{s_Y/\sqrt{n}} = \frac{31-30}{8/\sqrt{100}} = 1.25$
- Calculate the p-value: $1 - \Phi(t^{act}) = 1 - \Phi(1.25) = 1 - 0.8944 = 0.1056$
- Or equivalently, 1.25 < 1.64 (5% ‘1-tailed’ critical value)
- Do not reject H_0 at the 5% level.

Confidence Intervals

- Useful when goal is to estimate a population parameter because it provides an indication of how variable the estimate is.

- A confidence interval is an interval which contains the true value of a parameter with a certain prespecified probability.

  - E.g. A 95% **confidence interval** for μ_Y is an interval that contains the true value of μ_Y in 95% of repeated samples.

Confidence Interval for the Population Mean

- when n is large,

$$\bar{Y} \ is \ approximately \ N(\mu_Y, \frac{\sigma^2{}_Y}{n})$$

- So the probability that $\bar{Y}$ will be within 1.96 standard deviations of the population mean μ_Y is 0.95.
- To say that $\bar{Y}$ lies within 1.96 standard deviations of μ_Y is to say that μ_Y is within 1.96 standard deviations of $\bar{Y}$
- So 95% of all samples will capture the true μ_Y in the interval from $\bar{Y} - 1.96\sigma_{\bar{Y}}$ to $\bar{Y} + 1.96\sigma_{\bar{Y}}$
- So a 95% confidence interval for μ_Y is

$$\bar{Y} - 1.96\sigma_{\bar{Y}} \leq \mu_Y \leq \bar{Y} + 1.96\sigma_{\bar{Y}} \quad or$$

$$\bar{Y} - 1.96\,\frac{\sigma_Y}{\sqrt{n}} \leq \mu_Y \leq \bar{Y} + 1.96\,\frac{\sigma_Y}{\sqrt{n}}$$

- A 95% confidence interval for μ_Y is

$$\bar{Y} - 1.96\frac{\sigma_Y}{\sqrt{n}} \leq \mu_Y \leq \bar{Y} + 1.96\frac{\sigma_Y}{\sqrt{n}}$$

- In practice, $\sigma_Y/\sqrt{n}$ is unknown – it must be estimated by $s_Y/\sqrt{n}$, the **standard error** of $\bar{Y}$.

- So the 95% confidence interval for μ_Y is

$$\bar{Y} - 1.96SE(\bar{Y}) \leq \mu_Y \leq \bar{Y} + 1.96SE(\bar{Y})$$

Confidence Intervals

- We can specify the confidence level we like

$$95\% \text{ confidence interval for } \mu_Y = \{\bar{Y} \pm 1.96 SE(\bar{Y})\}$$

$$90\% \text{ confidence interval for } \mu_Y = \{\bar{Y} \pm 1.64 SE(\bar{Y})\}$$

$$99\% \text{ confidence interval for } \mu_Y = \{\bar{Y} \pm 2.58 SE(\bar{Y})\}$$

Confidence Intervals: An E.g.

You randomly sample 100 Singaporean workers and find that

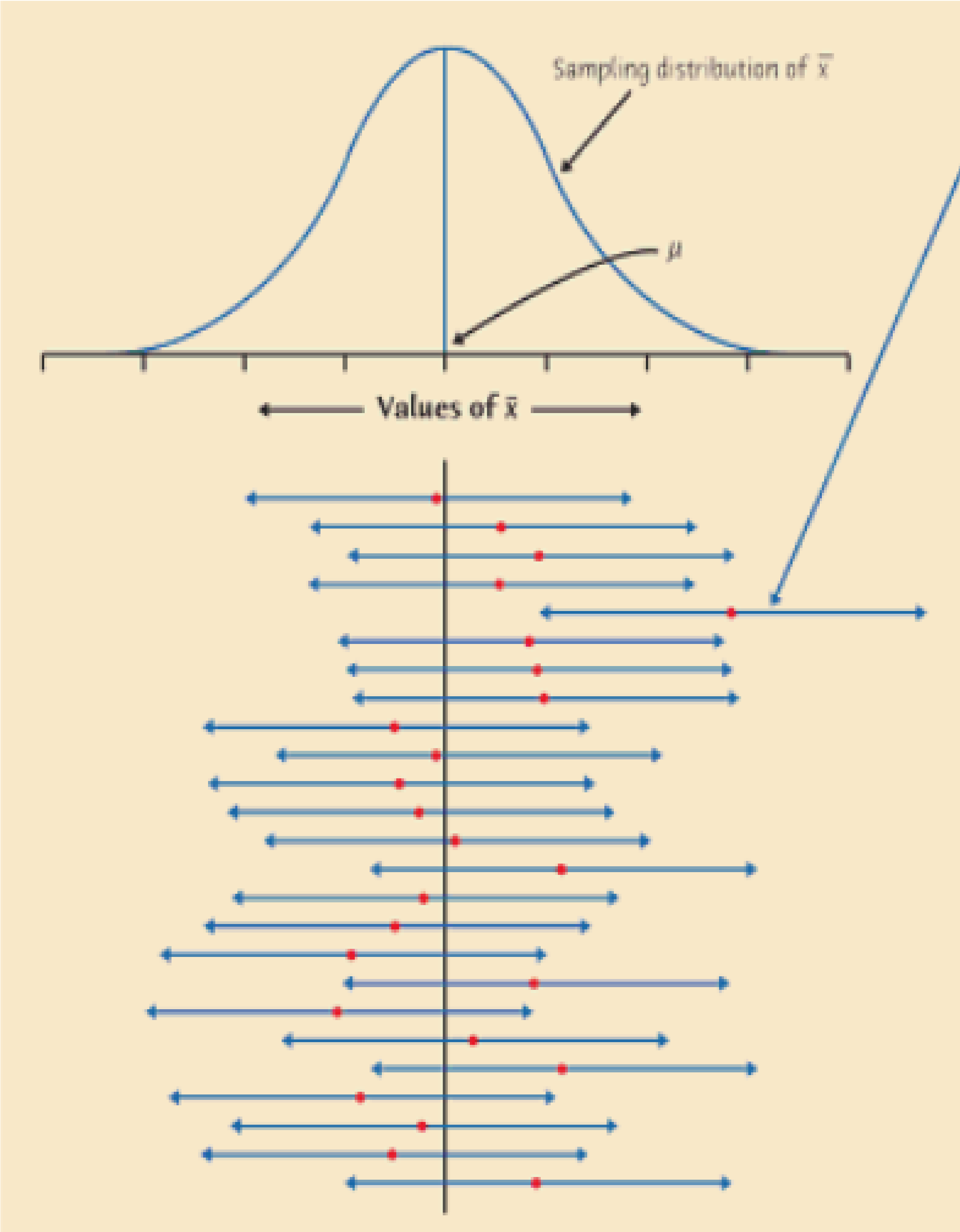
$$\bar{Y} = \$31, \qquad s_Y = 8, \qquad n = 100$$

- The 95% confidence interval for the mean hourly earnings of Singaporeans is

$$\mu_Y = \{\bar{Y} \pm 1.96 SE(\bar{Y})\} = \left\{\bar{Y} \pm 1.96\frac{S_Y}{\sqrt{n}}\right\} = \left\{31 \pm 1.96\frac{8}{\sqrt{100}}\right\} =$$

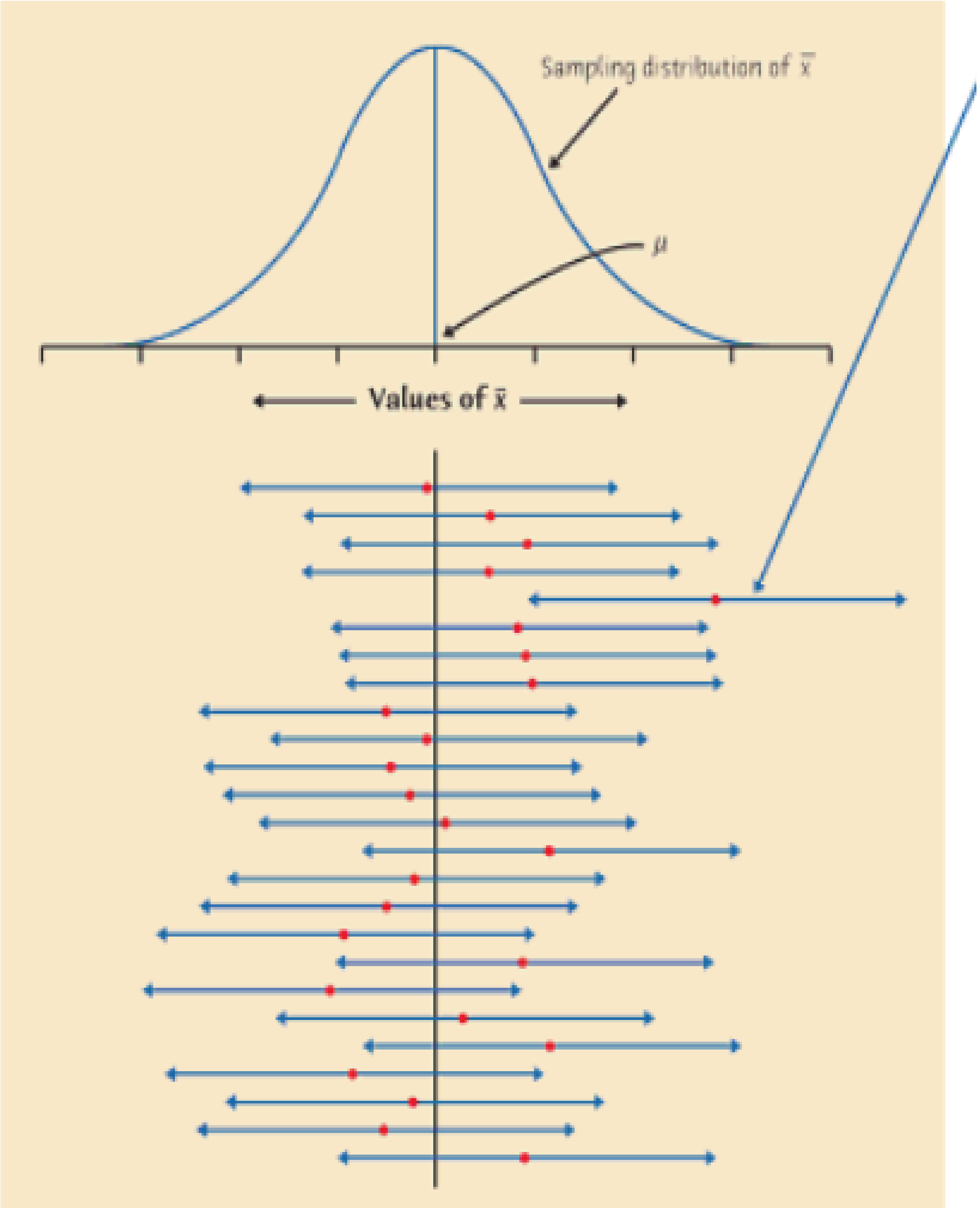
[$29.43, $32.57]

- What this says:
  - We are 95% confident that the mean earnings of Singaporeans lies between $29.43 and $32.57 per hour.
- What this does not say:
  - The probability is 95% that the true mean earnings falls between $29.43 and $32.57 per hour.



This interval misses the true μ. The others all capture μ.

- The figure illustrates the behavior of 95% CI in repeated sampling.
- Here, there are 25 samples, giving these 95% CIs.
- The centre of each interval is at $\bar{X}$ and so varies from sample to sample. The "margin of error", $\pm \mathbf{1.96SE}(\bar{\boldsymbol{X}})$, is the same for each interval.
- In the long run, ***95% of all samples give an interval that contains the true μ***
- We are not sure if our sample is one of the 95% where the interval contains μ or one of the "unlucky" 5%
- So we say we are 95% confident that the true μ lies between the interval $29.43 and $32.57



- This is ***not the same*** as saying “the probability is 95% that the true μ falls between \$29.43 and \$32.57”.
- The true μ, either is, or is not, between \$29.43 and \$32.57.