

STA 291

Summer 2008

Lecture 7

The Normal Probability Distribution

- Carl Friedrich Gauß (1777-1855), ***Gaussian Distribution***
- Perfectly ***symmetric*** and ***bell-shaped***
- Two parameters:
mean μ and ***standard deviation s***
- The ***68%-95%-99.7%*** rule applies to the normal distribution
- Actually, to be precise, the more correct percentages are 68.26895% - 95.44997% - 99.73002%
- That is, the probability concentrated within 1 standard deviation of the mean is always 0.6826895 for a normal distribution

More Examples for Finding z-Values

- Verify that the z-value for a right-hand tail probability
 - of 0.05 is $z=1.65$
 - of 0.01 is $z=2.33$
 - of 0.5 is $z=???$

Finding z-Values for Percentiles

- For a normal distribution, how many standard deviations from the mean is the 90th percentile?
- Or: What is the value of z such that 0.90 probability is less than $\mu + z s$?
- If 0.9 probability is less than $\mu + z s$, then there is 0.4 probability between 0 and $\mu + z s$ (because there is 0.5 probability less than 0)
- $z=1.28$
- The 90th percentile of a normal distribution is 1.28 standard deviations above the mean

Finding z-Values for Two-Tail Probabilities

- What is the z-value such that the probability is 0.1 that a normally distributed random variable falls more than z standard deviations **above or below** the mean
- Symmetry: we need to find the z-value such that the right-tail probability is 0.05 (more than z standard deviations **above** the mean)
- $z=1.65$
- 10% probability for a normally distributed random variable is outside 1.65 standard deviations from the mean, and 90% is within 1.65 standard deviations from the mean
- Find the z-value such that the probability is 0.5 that a normally distributed random variable falls more than z standard deviations **above or below** the mean

Quartiles of Normal Distributions

- Median: $z=0$
(0 standard deviations above the mean)
- Upper Quartile: $z = 0.67$
(0.67 standard deviations above the mean)
- Lower Quartile: $z = - 0.67$
(0.67 standard deviations below the mean)

Example

- Suppose that the weekly use of gasoline for motor travel by adults in North America is approximately normally distributed with mean 16 gallons and standard deviation 5 gallons.
- What proportion of adults use more than 20 gallons per week?
- Find the lower and upper quartiles and interquartile range of gasoline use.

z-Scores

- The z-score for a value x of a random variable is the number of standard deviations that x is above μ
- If x is below μ , then the z-score is negative
- The z-score is used to compare values from different normal distributions

Calculating z-Scores

- You need to know x , μ , and s to calculate z

$$z = \frac{x - m}{s}$$

Tail Probabilities

- SAT Scores: Mean=500,
Standard Deviation =100
- The SAT score 700 has a z-score of $z=2$
- The probability that a score is **beyond** 700 is the tail probability of $z=2$
- Table 3 provides a probability of 0.4772 between mean=500 and 700.
- Therefore, the right-tail probability for a score of 700 equals 0.0228
- 2.28% of the SAT scores are **beyond** 700 (**above** 700)

Tail Probabilities

- SAT score 450 has a z-score of $z=-0.5$
- The probability that a score is **beyond** 450 is the tail probability of $z=-0.5$
- Because of the symmetry of the normal distribution, there is exactly as much probability beyond -0.5 to the left as beyond 0.5 to the right
- Table 3 provides a probability of 0.1915, corresponding to a tail probability of 0.3085 for $z=0.5$
- 30.85% of the SAT scores are **beyond** 450 (**below** 450)

z-Scores

- The z-score is used to compare values from different normal distributions
- SAT: $\mu=500$, $s=100$
- ACT: $\mu=18$, $s=6$
- What is better, 650 in the SAT or 25 in the ACT?

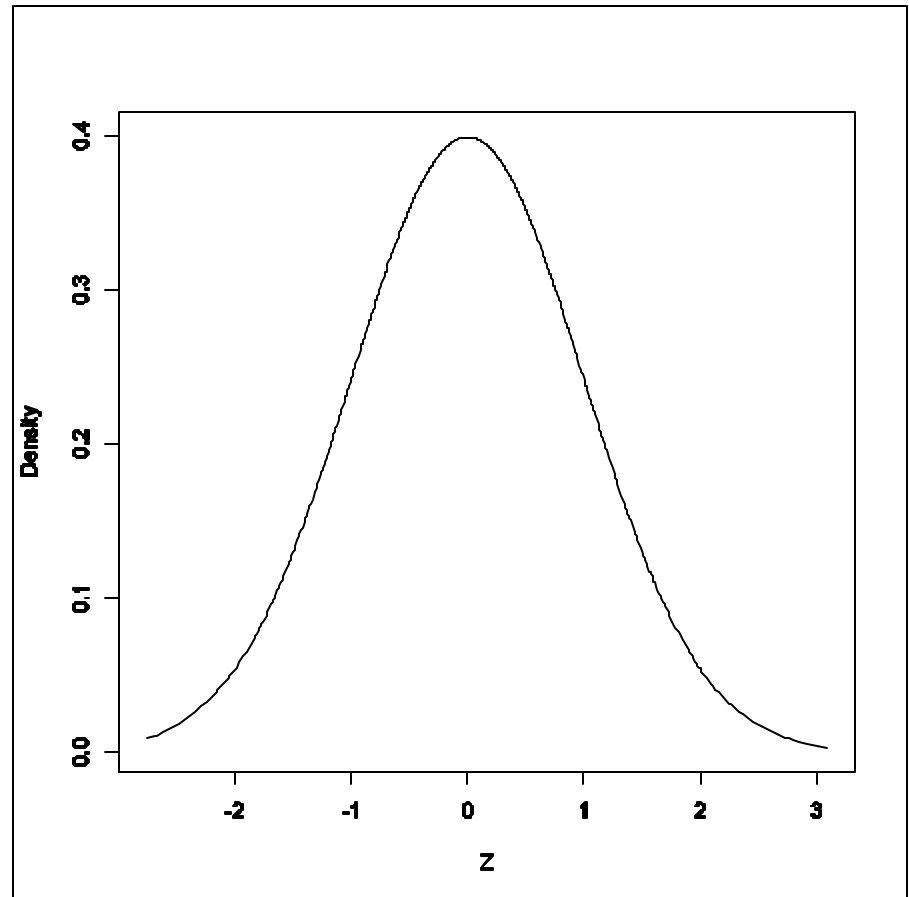
$$z_{SAT} = \frac{x - \mathbf{m}}{\mathbf{s}} = \frac{650 - 500}{100} = 1.5$$

$$z_{ACT} = \frac{x - \mathbf{m}}{\mathbf{s}} = \frac{25 - 18}{6} = 1.17$$

Corresponding tail probabilities?
How many percent have better
SAT or ACT scores?

Standard Normal Distribution

- The standard normal distribution is the normal distribution with mean $\mu=0$ and standard deviation $s=1$



Standard Normal Distribution

- When values from an arbitrary normal distribution are converted to z-scores, then they have a standard normal distribution
- The conversion is done by subtracting the mean μ , and then dividing by the standard deviation s

$$z = \frac{x - \mathbf{m}}{\mathbf{S}}$$

Example

- The scores on the Psychomotor Development Index (PDI) are approximately normally distributed with mean 100 and standard deviation 15. An infant is selected at random.
- Find the probability that the infant's PDI score is at least 100.
- Find the probability that PDI is between 97 and 103.
- Find the z-score for a PDI value of 90. Would you be surprised to observe a value of 90?
- Suppose we convert all the PDI observations to z-scores; that is, for each infant, subtract 100 from the value of PDI and divide by 15. Then, what is the distribution of the z-scores called? What are the mean and standard deviation of these z-scores?

Typical Questions

- One of the following three is given, and you are supposed to calculate one of the remaining
 1. Probability or percentage (right-hand, left-hand, two-sided, middle)
 2. z-score
 3. Observation x
- In converting between 1 and 2, you need Table 3 or one of the online tools. Probabilities are in the body of the table.
- In transforming between 2 and 3, you need mean and standard deviation and one of the following formulas

$$z = \frac{x - \mathbf{m}}{\mathbf{S}}$$

$$x = \mathbf{m} + z\mathbf{S}$$

Note: Most of the time, mu and sigma are provided. If not, things can be a bit more tricky.

9. Sampling Distributions

- A sampling distribution is a probability distribution that determines probabilities of the possible values of a sample statistic, for example the sample mean
- If you repeatedly take random samples and calculate the sample mean each time, the distribution of the sample mean follows a pattern
- This pattern is the sampling distribution

Sampling Distribution: Example

- If we randomly choose a student from this class, then with about 0.5 probability, he/she is majoring in Arts&Sciences or Business&Economics
- We can take a random sample and find the sample proportion of AS/BE students
- Define a variable X where
 - $X=1$ if the student is in AS/BE
 - and $X=0$ otherwise

Sampling Distribution: Example (contd.)

- If we take a sample of size $n=4$, the following 16 samples are possible:

$(1, 1, 1, 1)$; $(1, 1, 1, 0)$; $(1, 1, 0, 1)$; $(1, 0, 1, 1)$;
 $(0, 1, 1, 1)$; $(1, 1, 0, 0)$; $(1, 0, 1, 0)$; $(1, 0, 0, 1)$;
 $(0, 1, 1, 0)$; $(0, 1, 0, 1)$; $(0, 0, 1, 1)$; $(1, 0, 0, 0)$;
 $(0, 1, 0, 0)$; $(0, 0, 1, 0)$; $(0, 0, 0, 1)$; $(0, 0, 0, 0)$

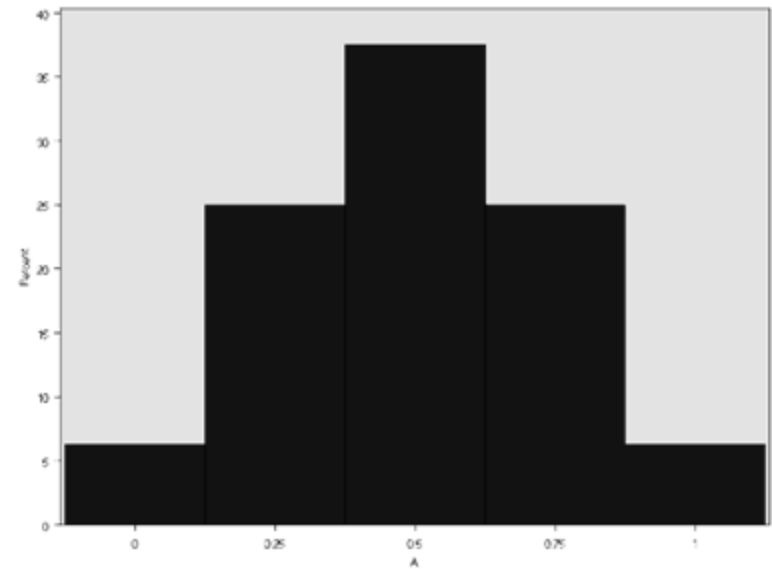
- Each of these 16 samples is equally likely because the probability of being in AS/BE is 50% in this class

Sampling Distribution: Example (contd.)

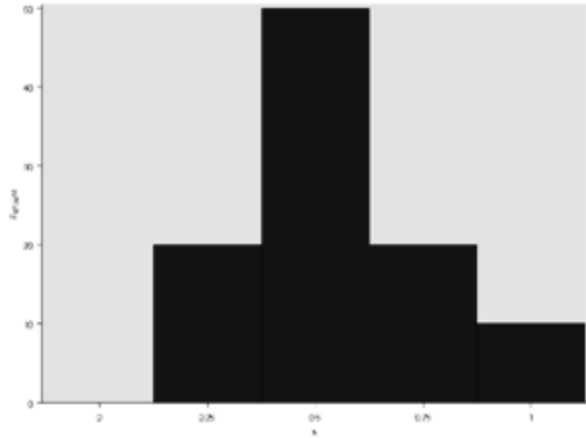
- We want to find the sampling distribution of the statistic “sample proportion of students in AS/BE”
- Note that the “sample proportion” is a special case of the “sample mean”
- The possible sample proportions are
 $0/4=0$, $1/4=0.25$, $2/4=0.5$, $3/4=0.75$, $4/4=1$
- How likely are these different proportions?
- This is the sampling distribution of the statistic “sample proportion”

Sampling Distribution: Example (contd.)

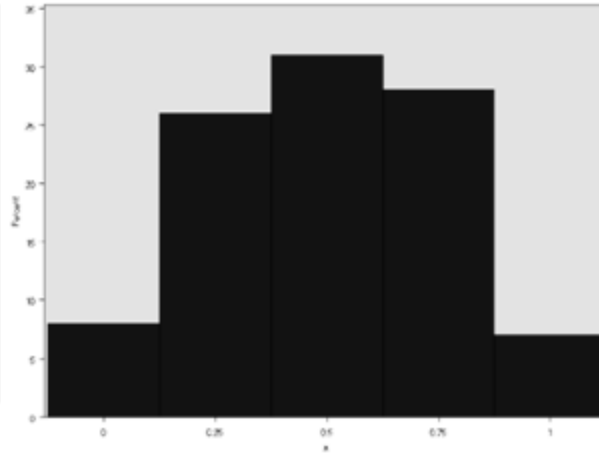
Sample Proportion of Students from AS/BE	Probability
0.00	$1/16=0.0625$
0.25	$4/16=0.25$
0.50	$6/16=0.375$
0.75	$4/16=0.25$
1.00	$1/16=0.0625$



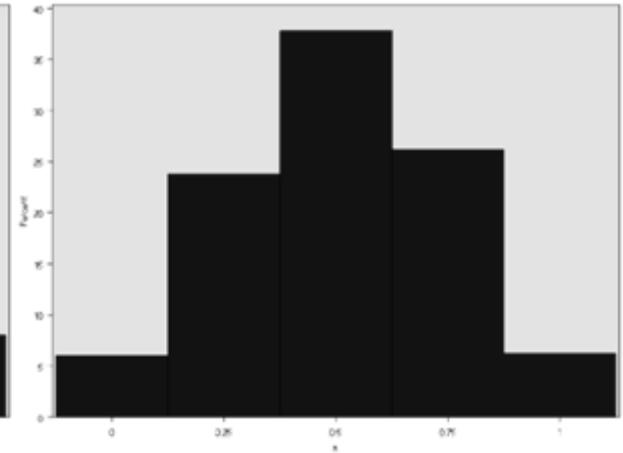
Sampling Distribution: Relative Frequencies of the different sample means (from Computer Simulations)



10 samples
of size $n=4$



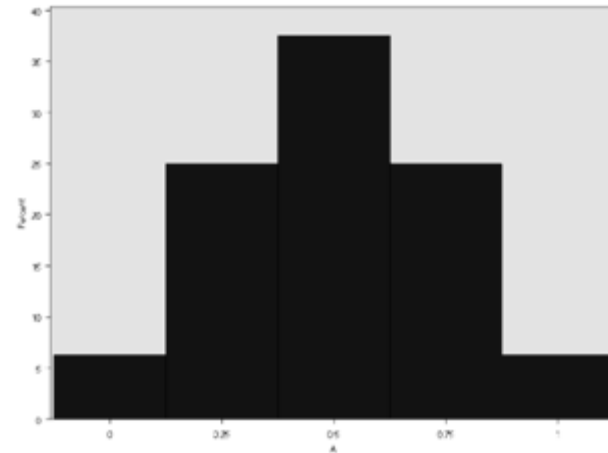
100 samples
of size $n=4$



1000 samples
of size $n=4$

Probability Distribution of the
Sample Mean

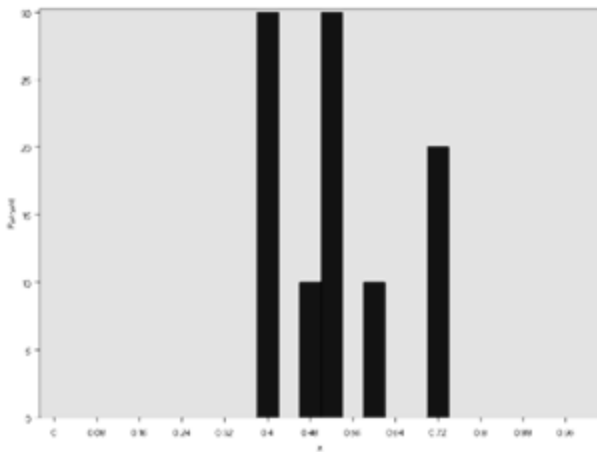
*This is what we should get if we
kept taking samples
infinitely often.*



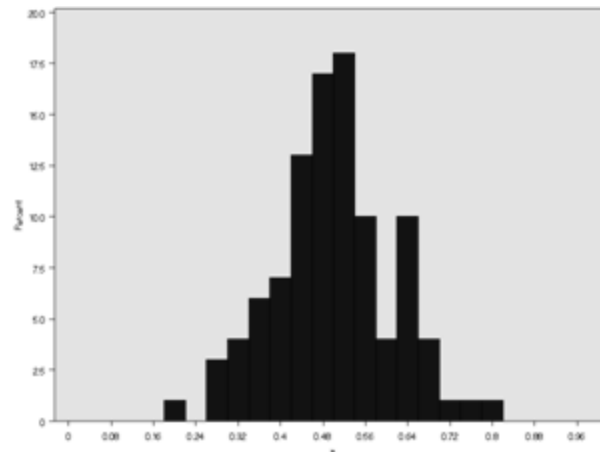
Reduce Sampling Variability

- Note: In the different simulations so far, we have only taken **more samples** of the same sample size $n=4$, we have not (yet) changed n .
- The larger the sample size, the smaller the sampling variability
- Let's simulate samples of size $n=25$ and see what happens...

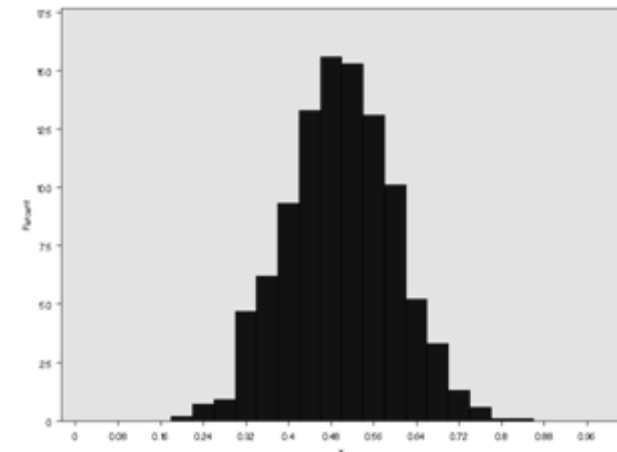
Sampling Distribution: Example (contd.), Computer Simulation



10 samples
of size $n=25$



100 samples
of size $n=25$



1000 samples
of size $n=25$

Interpretation

- If you take samples of size $n=4$, it may happen that nobody in the sample is in AS/BE
- If you take larger samples ($n=25$), it is highly unlikely that nobody in the sample is in AS/BE
- The sampling distribution is more concentrated around its mean
- The mean of the sampling distribution is the population mean: In this case, it is 0.5

Using the Sampling Distribution

- In practice, you only take **one** sample
- The knowledge about the sampling distribution helps to determine whether the result from the sample is reasonable given the model
- For example, our model was
 $P(\text{randomly selected student is in AS/BE})=0.5$
- If the sample mean is very unreasonable given the model, then the model is probably wrong

Effect of Sample Size

- The larger the sample size n , the smaller the standard deviation of the sampling distribution for the sample mean

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

- Larger sample size = better precision
- As the sample size grows, the sampling distribution of the sample mean approaches a normal distribution
 - Usually, for about $n=30$, the sampling distribution of the sample mean is close to normal (Gaussian)
 - This is called the “Central Limit Theorem”

Sampling Distribution of the Sample Mean

- When we calculate the sample mean \bar{X} , we do not know how close it is to the population mean μ
- μ is unknown
- The sampling distribution tells us with which probability the sample mean falls within, say, 3 units of μ

Parameters of the Sampling Distribution

- If we take random samples of size n from a population with population mean μ and population standard deviation σ , then the sampling distribution of \bar{X}
 - has mean μ
 - and standard error $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
- The standard deviation of the sampling distribution of the mean is sometimes called “standard error” to distinguish it from the population standard deviation

Standard Error

- Intuitively, larger samples yield more precise estimates
- Example:
 - $X=1$ if student is in AS/BE, $X=0$ otherwise
 - The population distribution of X has mean $p=0.5$ and standard deviation

$$\mathbf{s} = \sqrt{p(1-p)} = 0.5$$

Standard Error

- Example (contd.):

- For a sample of size $n=4$, the standard error of \bar{X} is

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{0.5}{\sqrt{4}} = 0.25$$

- For a sample of size $n=25$,

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{0.5}{\sqrt{25}} = 0.1$$

- Because of the approximately normal shape of the distribution, we would expect \bar{X} to be within 3 standard errors of the population mean (with 99.7% probability)

Central Limit Theorem

- For random sampling, as the sample size n grows, the sampling distribution of the sample mean \bar{X} approaches a normal distribution
- Amazing: This is the case even if the population distribution is discrete or highly skewed
- The Central Limit Theorem can be proved mathematically
- We can verify it experimentally in the lab sessions

Central Limit Theorem

- Usually, the sampling distribution of \bar{X} is approximately normal for $n=20$ or 30
- In addition, we know that the parameters of the sampling distribution are \mathbf{m} and $\mathbf{s}_{\bar{X}} = \frac{\mathbf{S}}{\sqrt{n}}$
- For example:

If the sample size is $n=25$, then with 95% probability, the sample mean falls between

$$\mathbf{m} - 1.96 \frac{\mathbf{S}}{\sqrt{n}} = \mathbf{m} - \frac{1.96}{5} \mathbf{S} \approx \mathbf{m} - 0.4\mathbf{S}$$

$$\text{and } \mathbf{m} + 1.96 \frac{\mathbf{S}}{\sqrt{n}} = \mathbf{m} + \frac{1.96}{5} \mathbf{S} \approx \mathbf{m} + 0.4\mathbf{S}$$

(\mathbf{m} = population mean, \mathbf{S} = population standard deviation)

Example

- Recall:
 - The scores on the Psychomotor Development Index (PDI) are approximately normally distributed with mean 100 and standard deviation 15. An infant is selected at random.
 - Find the probability that the infant's PDI score is at least 100.
 - Answer: 0.5
 - Find the probability that PDI is between 97 and 103.
 - Answer: 0.16
 - Find the z-score for a PDI value of 90. Would you be surprised to observe a value of 90?
 - Answer: -0.67; no, not surprised because 25% of the observations would even be below 90

Revised Example

- Refer to the previous exercise. A study uses a random sample of 225 infants
- Describe the sampling distribution of the sample mean PDI
- Find the probability that the sample mean falls between 97 and 103
- Find the z-score from the sampling distribution corresponding to a sample mean of 90 when the sample size is 225. Would you be surprised to observe a sample mean PDI of 90?
- Compare the results with those on the previous slide, and interpret the differences

Multiple Choice Question 1

The standard error of a statistic describes

1. The standard deviation of the sampling distribution of that statistic
2. The standard deviation of the sample measurements
3. How close that statistic is likely to fall to the parameter that it estimates
4. The variability in the values of the statistic for repeated random samples of size n
5. The error that occurs due to nonresponse and measurement errors

Multiple Choice Question 2

The Central Limit Theorem implies that

1. All variables have approximately bell-shaped sample distributions if a random sample contains at least 30 observations
2. Population distributions are normal whenever the population size is large
3. For large random samples, the sampling distribution of the sample mean (\bar{X}) is approximately normal, regardless of the shape of the population distribution
4. The sampling distribution looks more like the population distribution as the sample size increases
5. All of the above