Probability and Statistics Guide

1 Introduction to Statistics

Statistics: is the science of collecting and analyzing data. In Statistics mathematical computations are used to support conclusions from a data.

Individuals: are the people or objects included in a statistical study. **A variable:** is a characteristic of the individuals to be measured.

Sample data: a portion of the data or data for only some of the individuals of interest.

Population data: the whole extension of the data we will like to analyze. Data from all individuals of interest.

A sample statistic: or sample estimate is a numerical attribute of a sample of data. It depends on the sample that you are considering.

A population parameter: a numerical attribute of a the whole **population**. It does not change when you consider different samples.

Levels of measurement: besides dividing the data in qualitative and quantitative, we have four levels of measurements indicating what kind of arithmetic is appropriate for the data:
Nominal or Categorical: Data that can not be ordered, like labels, names or categories.
Ordinal: Data can be ordered, but differences between the data are meaningless.
Interval: Data can be ordered and differences and averages are meaningful.
Ratio: Data can be arranged in order, addition, differences and also ratios are meaningful.

Simple random Sample of n measurements: A selection of a subset of **n** elements or individuals of the population, when all members of the population have the same chance of being selected and every sample of the given size **n** has the same chance of being selected. The number **n** is called the sample size.

Descriptive Statistics: describe quantitatively features of a sample data. It aims to summarize or describe the sample using **statistics**. Descriptive Statistics can be univariate when studies features of just one variable or multivariate when aims to relate features of two or more variables.

Inferential Statistics: aims to use the data to learn about the whole population. It is based heavily on the theory of Probability.

QUESTIONS FROM SECTION 1

1. Classify each of the following data according to the *level of measurement* (that is state whether it is nominal, ordinal, interval, or ratio):
   (a) The telephone numbers in a telephone directory.
   (b) The scores of a class in an exam.
   (c) Absolute temperatures (that is temperatures measured in Kelvin degrees).
   (d) Motion Picture Association of America ratings description (G, PG, PG-13, R, NC-17).
   (e) Average monthly precipitation in inches for New York, NY.
   (f) Average monthly temperature (in degrees Fahrenheit) for New York, NY.
2. What is the difference between a sample statistic and a population parameter?

3. Explain in your own words what we understand by simple random sample of a population?

4. Why do you think that the technique of simple random sampling is difficult to use in practice?

2 Organizing Data

A frequency table: partitions the data into classes of equal width. **The class width:** is the smallest integer greater or equal to

$$\frac{Largest\,value - Smallest\,value}{Number\,of\,classes}.$$

The lower class limit: is the smallest value within a class. **The upper class limit:** is the highest data value that can fit in a class. **The class width:** is the difference between **The upper class limit** and **The lower class limit**. To determine **the class boundaries** you subtract .5 from the lower class limit and add .5 to the upper class limit.

QUESTIONS FROM SECTION 2

1. A group of 25 people were observed regarding their TV habits and were found to spend the following number of hours per week watching television:

30	32	34	36	36
37	39	39	41	41
42	42	43	43	44
45	45	45	46	47
47	49	49	52	53

In order to display the data in clearer form,

(a) determine the class width for four (4) classes,

(b) construct a frequency distribution showing the class limits for the four classes,

(c) in the table, show the class boundaries and the class marks,

(d) construct a histogram, labeling the class boundaries. Is the graph symmetrical, skewed left or skewed right?

2. The following data represents the outcome of a scientific study:

15	16	18	18	22
27	28	29	29	30
32	32	33	33	34
35	35	35	36	38

In order to display the data in clearer form,

(a) determine the class width for three (3) classes,

(b) construct a frequency distribution showing the class limits for the four classes,

(c) in the table, show the class boundaries and the class marks,

(d) construct a histogram, labeling the class boundaries. Is the graph symmetrical, skewed left or skewed right?

3 Averages and variations

Measures of central tendency: are different ways to indicate the typical or central value in a distribution of data. There are three main measures: the mode, the median and the mean.

The mode: is the single data that occurs most frequently.
The median: is the middle value of the data once the data has been arranged in order.
The mean: is the average, that is, the sum of all data values devided by the number of data values.

Measures of variation: are measures of the dispersion of the data.

The variance: σ^2 is " the average squared deviation from the mean". We use square to prevent it from being zero. To find **the standard deviation** σ we use square root to go back to the original units of measurements. In the sample standard deviation to be able to use s as an "unbiased" estimation of σ, the sum of the squares of the deviations is divided by one less than the sample size.

Parameter	**Defining formula**	**Computational formula**
Population mean	$\mu = \frac{\sum x}{N}$	$\mu = \frac{\sum x}{N}$
Population standard deviation	$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$	$\sigma = \sqrt{\frac{\sum x^2 - (\sum x)^2/N}{N}}$

statistic	**Defining formula**	**Formula**
sample mean	$\bar{x} = \frac{\sum x}{n}$	$\bar{x} = \frac{\sum x}{n}$
sample standard deviation	$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$	$s = \sqrt{\frac{\sum x^2 - (\sum x)^2/n}{n-1}}$

The coefficient of variation: measures the spread of the data relative to the mean and is given by the formula:

$$C.V. = \frac{s}{\bar{x}} \times 100\%$$

The range: is the simplest measure of variation, computed as the highest value minus the lowest value.

The quartiles: Divide the data in four equal parts. **The interquartile range IQR:** is the difference $Q_3 - Q_1$ between the third and first quartiles. It defines how spread out is the center 50 % of the data.

The p-th percentile of a distribution of data is a value such that $p\%$ of the data fall at or below it $(100 - p)\%$ of the data fall at or above it.

Outliers: are values that are so low or so high that they seem to stand apart from the rest of the data. Outliers may represent data collection errors or data entry errors.

Potential outliers: can be detected using Quartiles Q_1, Q_3 as values outside the range

$$[Q_1 - 1.5(IQR), Q_3 + 1.5(IQR)].$$

Potential outliers: can be detected using standard deviation s as values outside the range

$$[\bar{x} - 2.5(s), \bar{x} + 2.5(s)].$$

QUESTIONS FROM SECTION 3

1. Calculate the range, mean, median, first and third quartiles, interquartile range, mode, variance, and standard deviation for the following population data.

   47 59 50 56 56 51 53 57 52 49

2. Find the mean, the range, and the standard deviation for the following set of sample data.

   10 9 12 11 8 15 9 7 8 6

3. Determine the range, median, mean and the sample standard deviation of the following data:

x	f
10.3	7
22	12
38.5	5
43.2	2

4. A consumer testing service obtained the following mileage (in miles per gallon) in five test runs for three different types of compact cars:

	First	**Second**	**Third**	**Fourth**	**Fifth**
	Run	**Run**	**Run**	**Run**	**Run**
Car A	28	32	28	34	30
Car B	31	31	29	29	31
Car C	32	29	28	32	30

   (a) If the manufacturer of Car A wants to advertise that their car performed the best in this test, which measure of central tendency (mean, median or mode) should be used to support their claim?

   (b) Which measure should the manufacturer of Car B use to claim that their car performed best, mean median or mode?

   (c) Which measure should the manufacturer of Car C use to support a similar claim?

5. In a class of 40 students, the grade of a particular student is the 90-th percentile. How many students score similar or more? Can she be sure that she passed the class?

6. In a set of data, what percent of the data is between Q_1 and Q_2 approximately?

7. Given the set of data:

1	20	21	24	26	26	26	27	28
32	33	33	34	36	39	43	43	47

   (a) Find the quartiles Q_1, Q_2 and Q_3, as well as the interquartile range $Q_3 - Q_1$.

   (b) Use quartiles to identify potential outliers.

4 Chebyshev's Theorem

Chebyshev's Theorem: For any set of data and for any constant k greater that 1 (not necessarily a whole number), the proportion of the data that falls within k standard deviations of the mean on either side is at least

$$1 - \frac{1}{k^2}.$$

Therefore using the values of $k = 2, 3, 4$ we obtain that for any set of data:

At least 75 % of the data lies in the interval from	$\mu - 2\sigma$ to $\mu + 2\sigma$
At least 88.9 % of the data lies in the interval from	$\mu - 3\sigma$ to $\mu + 3\sigma$
At least 93.8 % of the data lies in the interval from	$\mu - 4\sigma$ to $\mu + 4\sigma$

QUESTIONS FROM SECTION 4

1. The mean value of the scores in a Statistics exam was 85 with a standard deviation of 4. Find an interval that contains at least 75% of the scores in that exam.

2. Florida's age distribution has mean value $\mu = 39.2$ and standard deviation $\sigma = 24.8$ (measured in years). Use Chebyshev's theorem to find an interval such that

   (a) the age in years of at least 75% of Florida's population is contained within that interval,

   (b) the age in years of at least 88.9% of Florida's population is contained within that interval,

   (c) the age in years of at least 93.8% of Florida's population is contained within that interval.

3. What value of the constant $k > 1$ we need to use to obtain a Chebyshev's interval with at least 50% of the data?

4. What value of the constant $k > 1$ we need to use to obtain a Chebyshev's interval with at least two thirds (2/3) of the data?

5. (For students with an Integration theory and Probability background) Prove the Chebyshev's theorem using integration on the set $|X - \mu| > k\sigma$ with respect to a probability measure dP.

5 Correlation and regression

A scatter diagram: is a graph in which data pairs are plotted as individual points in a system of Cartesian coordinates. The variable x is called the explanatory or independent variable and the variable y is the response variable or dependent variable.

A linear regression: Finds a model of the response variable y as a linear function of the independent variable x.

High or Strong correlation: when the points are close to a straight line.

Positive linear correlation: The variables x and y are said to have positive linear correlations if low values of x are associated to low values of y and high values of x correspond to high values of y.

Negative linear correlation: The variables x and y are said to have negative linear correlations if low values of x are associated to high values of y and high values of x correspond to low values of y.

The correlation coefficient or **Pearson's correlation coefficient** r**:** is a numerical measure of the linear relation between two variables. It is always a number $-1 \leq r \leq 1$ and it admits a geometric interpretation as the cosine of the angle between the vectors $x - \bar{x}$ and $y - \bar{y}$. The $r = 1$ indicates **perfect positive correlation** (the points on the plot lie on a line of positive slope) and $r = -1$ is an indication of **perfect negative correlation** (points (x, y) are on a line of negative slope). On the other hand $r \approx 0$ will be an indication of **little or no linear correlation** whatsoever between x and y.

statistic	**Defining formula**	**Computational formula**
Coefficient of linear correlation r	$r = \dfrac{1}{n-1}\sum \dfrac{(x-\bar{x})}{s_x}\dfrac{(y-\bar{y})}{s_y}$	$r = \dfrac{n\sum xy - (\sum x)(\sum(y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$

Least squares line: the line such that the sum of squares of the difference of the y-values between the line and the points is as small as possible.

Fact: The least squares lines of equation $y = bx + a$ may or may not pass by any of the points, but it always contains the point

$$(\bar{x}, \bar{y}) = \left(\frac{\sum x}{n}, \frac{\sum y}{n}\right).$$

On the other hand the slope b is giving by the formula:

$$b = \frac{n\sum xy - (\sum x)(\sum(y)}{n\sum x^2 - (\sum x)^2},$$

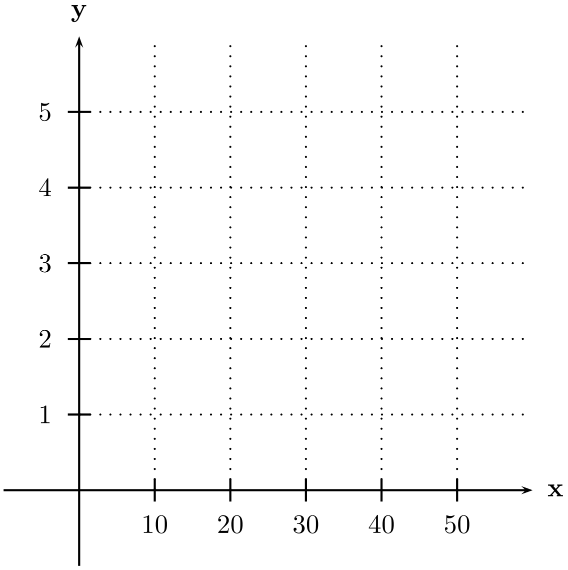
and we have an equation for the least squares line of the form $y - \bar{y} = b(x - \bar{x})$.

The coefficient of determination: is the square r^2 of the coefficient of correlation r. It reflects what portion of the variance of the response variable y can be explained by the variance of the independent variable x and the model $\hat{y} = a + bx$. The proportion $1 - r^2$ of the variance cannot be explained using the model.

QUESTIONS FROM SECTION 5

1. The manager of a salmon cannery suspects that the demand for her product is closely related to the disposable income of her target region. To test out this hypothesis she collected the following data for five different target regions, where x represents the annual disposable income for a region in millions of dollars and y represents sales volume in thousands of cases.

x	y
10	1
20	3
40	4
50	5
30	2

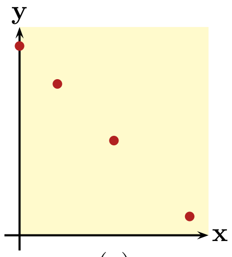


   (a) Draw the scatter graph of this set of data.
   (b) Compute the correlation coefficient r.
   (c) Compute the coefficient of determination r^2.

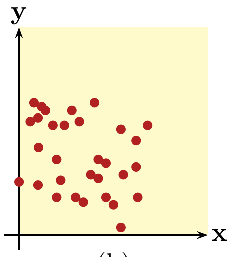
(d) Find and graph the least square line.

(e) If a region has disposable annual income $\$25,000,000$ what is the predicted sales volume?

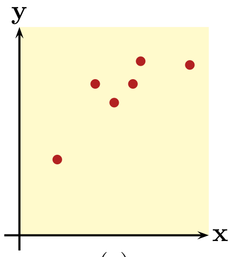
2. Match the appropriate statement about r and the scatter diagrams.



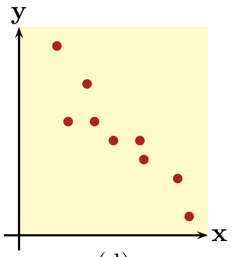
(a)



(b)



(c)

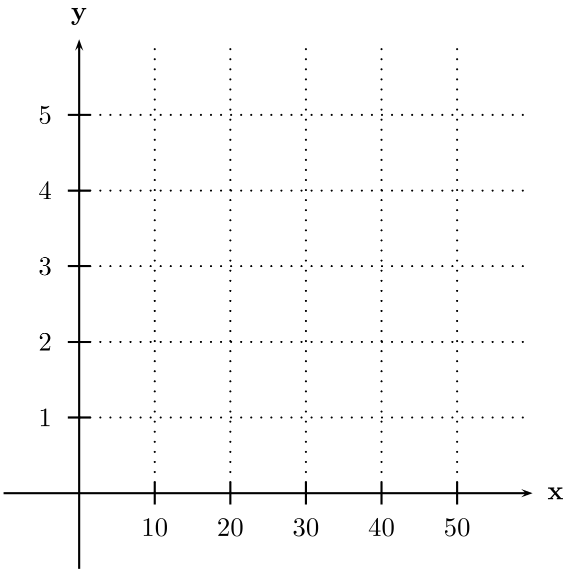


(d)

A. $-1 < r < 0$ B. $r = 0$. C. $r = -1$. D. $0 < r < 1$

3. The following table represents two sets of data:

x	y
3	4.2
5	4
12	3.5
17	3.8
23	2.4
48	.5



(a) Draw the scatter graph of this set of data.

(b) Based on the graph do you expect the correlation coefficient to be positive, negative or close to zero?

(c) Compute the coefficient of linear correlation r.

(d) Compute the coefficient of determination r^2.

(e) Find and graph the least square line.

(f) What will be the y predicted by the model for $x = 30$?

4. (It requires Calculus) Show that the least squares line always contains the point $(\bar{x}, \bar{y})$.

6 Introduction to probability theory

A statistical experiment: is any random activity that results in a definite outcome.

An event: is a set of one or more outcomes of a statistical experiment or observation. **A simple event:** is one particular outcome of a statistical experiment.

The sample space: is the set Ω of all simple events. The set of events $\mathcal{F}$ is a collection of subsets of Ω.

Probability: is a numerical measure, denoted $P(A)$, between 0 and 1 that describes the likelihood that an event A will occur. The higher the probability of an event, the more certain that the event will occur. If $P(A) = 1$, the event A is certain to occur and if $P(A)$=0, the event A is certain not to occur (impossible).

The complement of the event A: is the event that A will not occur. It is denoted by A^c

Mutually exclusive events: Two events are mutually exclusive if they **cannot** occur together. That is when

$$P(A\,and\,B) = 0.$$

Addition rule for mutually exclusive events: states that for A and B mutually exclusive

$$P(A \text{ or } B) = P(A) + P(B).$$

General addition rule: For any events (not necessarily mutually exclusive) we have:

$$P(A\,\mathrm{or}\,B) = P(A) + P(B) - P(A\,and\,B).$$

We choose the collection $\mathcal{F}$ in such a way that we can always take **complements** and **unions** (**or**). Also, the whole sample space Ω is always in $\mathcal{F}$. A collection $\mathcal{F}$ of sets with these properties is called a **"tribe"** and it represents the collection of measurable events.

A probability space: Is a triple $(\Omega, \mathcal{F}, P)$ consisting of a sample space Ω, a space of measurable events $\mathcal{F}$ and a probability assignment $P\colon \mathcal{F} \to [0, 1]$ in such a way that

1. The probability of the total space $P(\Omega) = 1$.
2. For any event A, the probability of the complement is $P(A^c) = 1 - P(A)$.
3. For any two mutually exclusive events A, B, $P(A\,\mathrm{or}\,B) = P(A) + P(B)$.

A probability assignment based on equally likely outcomes: uses the formula

$$P(A) = \frac{number\,of\,favorable\,outcomes}{total\,number\,of\,outcomes}.$$

Two events are independent: if the occurrence or nonoccurrence of one event does not change the probability that the other event will occur.

The multiplication rule for independent events: states that for A and B independent

$$P(A \text{ and } B) = P(A).P(B).$$

The conditional probability $P(A|B)$: denotes the probability that event A will occur given that event B already occurred. For events A, B that are not independent we have $P(A|B) \neq P(A)$ or $P(B|A) \neq P(B)$.

In general $P(A|B) \neq P(B|A)$.

The general multiplication rule: For any events A and B (not necessarily independent) we have:

$$P(A \text{ and } B) = P(A).P(B|A), \quad P(A \text{ and } B) = P(B).P(A|B).$$

The conditional probability $P(A|B)$: when $P(A) \neq 0$ can be found using the formula:

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}.$$

Baye's theorem: The conditional probability of an event can be expressed in terms of prior knowledge of conditions that may be related to the event:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}.$$

The complement of the union: can be found with the formula:

$$P(\text{neither } A \text{ nor } B) = 1 - P(A) - P(B) + P(A \text{ and } B).$$

Total probability formula: For a partition of the sample space in events $B_1, B_2, \ldots, B_n$, we have

$$P(A) = \sum_{i=1}^{n} P(A \text{ and } B_i) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n).$$

QUESTIONS FROM SECTION 6

1. Given $P(E^C) = 0.3$, $P(F) = 0.35$, and $P(F|E) = 0.25$ find
   (a) $P(E \text{ and } F)$
   (b) $P(E \text{ or } F)$
   (c) $P(E|F)$.
2. Two dice are rolled. Find the probability of the following events:
   (a) Both numbers are 6.
   (b) The first dice gives 5 and the second 6.
   (c) There is one 5 and one 6.
   (d) The sum is equal to 10.
   (e) Both are 6 or the sum 10.
   (f) The sum is more than 5 but less than 8.
   (g) Both numbers are even.
   (h) One number is even and one number is odd.
3. Calculate by hand (without a calculator). Show all work:
   (a) 5!
   (b) $C(12, 3)$.
   (c) $C(1,000, 100, 2)$.
4. An urn contains three yellow, four green, and five blue balls. Two balls are randomly drawn without replacement. Find the probability of the following events:
   (a) Both balls are blue.
   (b) The first ball is green and the second yellow.
   (c) There is one green and one yellow ball.

5. Repeat the previous exercise but now assume that the balls are drawn with replacement.

6. Three cards are randomly drawn from a standard 52 card deck without replacement. Find the probability of the following events:

   (a) All cards are red.

   (b) There are two red and one black card.

   (c) All cards are spades.

   (d) There is one spade, one club, and one diamond.

   (e) All cards are aces.

   (f) Two cards are aces and one card is a king.

7. Most of the time, a medical test is able to correctly indicate if a person has a condition. However, some of the time, there are false positives (it indicates the condition is present when it is not) or false negatives (it indicates the condition is not present when it is there). Use the table below to determine the probabilities for a randomly selected person from the population.

	condition present	condition not present	row total
Test Result +	125	10	135
Test Result −	15	50	65
column total	140	60	200

   (a) What is the probability of a false positive?

   (b) What is the probability of either a false positive or a false negative?

   (c) What is the probability of a positive test result given that the condition is present?

   (d) What is the probability that the condition is present given a positive test result?

   (e) What is the probability of either a negative test result or the condition is not present?

8. One college found that during one semester $1,259$ students in its four most popular majors had the following class distributions. Use the table below to determine the probabilities for a randomly selected student in this group.

	first year	sophomore	junior	senior	row total
Business	115	90	105	111	421
Psycology	88	95	91	96	370
Nursing	85	81	79	76	321
Biology	63	45	25	14	147
column total	351	311	300	297	1259

   (a) What is the probability of being a business major?

   (b) What is the probability of not being a biology major?

   (c) What is the probability of being a senior and majoring in psychology?

   (d) What is the probability of being a senior or sophomore?

   (e) What is the probability of being a senior or biology major?

   (f) What is the probability of being a junior, given being a nursing major?

   (g) What is the probability of being a nursing major, given being a junior?

9. An island is a habitat for 208 species of birds. 82 of these species are found only on this particular island. 75 species are seabirds. 12 are a species of seabird and are found only on this particular island. One species of bird is chosen at random.

   (a) What is the probability it is a seabird or unique to this island?

(b) What is the probability it is neither a seabird nor unique to this island?

10. (a) A company is looking hire more sales staff. The human resources department accepts only the 45% of the submitted resumes that meet the hiring criteria. The managers then select 20% of the applicants with accepted resumes to come in for an interview. What is the probability that an applicant selected at random will have her resume accepted and be granted an interview?

    (b) In one high school, the athletic director found that 4% of the varsity athletes had concussions while playing at the school and 18% had severe sprains and 1% had experienced both. What is the probability that a randomly selected varsity athlete has either had a concussion or a severe sprain?

7 Random Variables and discrete probability distributions.

A random variable: is a quantitative variable X that takes random outcomes. It can be though of as a function from the sample space to the real numbers, in such a way that we can always measure the probability of X being on a given interval.

A discrete random variable: can take at most countable many values.

A continuous random variable: takes all the values of an interval in the real line.

The probability distribution or density function ρ for a discrete random variable X: is an assignment of a probability to each value taken by a discrete random variable in such a way that **the sum of all probabilities is always 1**.

The expected value or mean of a discrete probability distribution is:

$$E(X) = \mu(X) = \sum xP(x).$$

The expected value is a linear operator:

$$E(aX + bY) = aE(X) + bE(Y).$$

The standard deviation and variance of a discrete probability distribution are:

$$\sigma(X) = \sqrt{\sum (x-\mu)^2 P(x)}, \qquad \sigma^2(X) = \sum (x-\mu)^2 P(x).$$

The variance satisfies the formula:

$$\sigma^2(X) = E(X^2) - E(X)^2.$$

And therefore:

$$\mu(ax + b) = a\mu(x) + b \qquad \sigma^2(aX + b) = a^2\sigma^2(X).$$

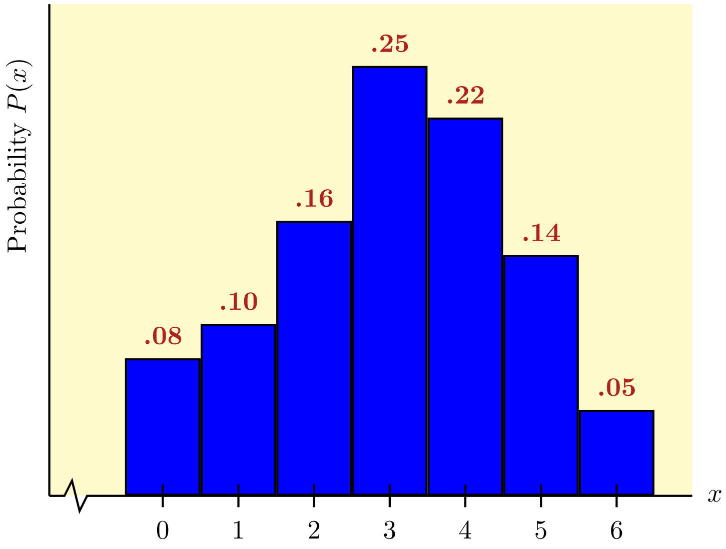
Some discrete probability distributions			
Distribution	Density function	mean	The variable X represents:
Poisson	$\rho(k) = P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$	$\mu = \lambda$	The number of events on an interval of fixed length.
Geometric (type I)	$\rho(k) = P(X = k) = (1-p)^{k-1}p$	$\mu = \frac{1}{p}$	The number of Bernoulli trials for the first success.
Binomial	$\rho(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$	$\mu = np$	The number of successes in n Bernoulli trials.

1. Complete the table in such a way that we have a discrete probability distribution.

x	2	3	4	5	6
$P(x)$	.25	.1	.3	.2	

   Sketch the graph of this distribution and calculate its expected value and standard deviation.

2. Find the expected value and the standard deviation of the probability distribution whose graph is shown:



3. A fair coin is tossed 7 times. Sketch the graph of the resulting binomial distribution.

4. (Bernoulli trials) Consider a random variable X with two positive outcomes, success (1) with probability p and failure (0) with probability $1 - p$. Find expected value and standard deviation for X.

5. (Requires Calculus) Prove that the formula for the distribution of Poisson is actually a distribution function. Prove that the mean of the Poisson distribution is exactly λ.

6. (Requires Calculus) Prove that the assignment $P(X = k) = \frac{1}{2^{k+1}}$ for $k = 0, 1, 2, \ldots$ represents a discrete probability distribution on the natural numbers. Find its mean and standard deviation.

8 The Binomial probability distribution

A binomial experiment: is an experiment with a fixed number **n** of independent trials, each of which can only have two possible outcomes (Bernoulli trials), and the probability of each outcome remains constant on each trial.

The probability of success: will be the probability **p** of one of the two outcomes on each trial. **The probability of failure:** will be the probability **q=1-p** of the other outcome.

Main question: What is the probability (for **r=0,1, . . . n**) of getting exactly **r** successful outcomes in **n** trials?

Answer: The probability of getting exactly r successes in n trials is

$$P(X=r) = C_{n,r}p^r(1-p)^{n-r} = \frac{n!}{r!(n-r)!}p^r(1-p)^{n-r}.$$

To probability $P(X=r)$ can be found using the Binomial Probability Distribution table.

As sum of independent Bernoulli events, the mean and standard deviation of a binomial probability distribution are:

$\mu = np$	$\sigma = \sqrt{np(1-p)}$

In the binomial distribution: The closer p is to .5 and the larger the number of sample observations n, the more symmetric the distribution becomes.

QUESTIONS FROM SECTION 8

1. Alice and Bob play the following game: two cards are randomly drawn (with replacement) from a standard 52-card deck, if they are both red Alice wins otherwise Bob wins. If they play these game 16 times what is the probability that Alice will win at most 4 times?

2. If 30% of the people in a community use the Library in one year, find the probability that in a random sample of 15 people
   (a) At most 7 use the Library,
   (b) Exactly 7 use the Library,
   (c) At least 5 use the Library,
   (d) No more than 2 use the Library,
   (e) Not less than 10 use the Library.

3. A basketball player makes 70% of the free throws he shoots. What is the probability that he will make more than 7 throws
   (a) If he tries 15 free throws?
   (b) If he tries 10 free throws?

4. Approximately 5% of the eggs in a store are cracked. Suppose you buy a dozen eggs from the store.
   (a) What is the probability that no more than one of your eggs is cracked?
   (b) What is the probability that fewer than 3 eggs are cracked?
   (c) Find the expected value and standard deviation of the number of cracked eggs.

5. A surgery has a success rate of 75%. Suppose that the surgery is performed on six patients. Find the expected value and the standard deviation of the number of successes.

6. One-third of all deaths are caused by heart attacks. If three deaths are chosen randomly, find the probability that none resulted from heart attack.

9 Continuous probability distributions

A continuous random variable X: takes all values on a whole interval of the real line.

The probability distribution ρ for a continuous random variable X: is an assignment of probability to each interval of the values taken by the variable X, in such a way that the total area under the curve given by

$$A = \int_{-\infty}^{\infty} \rho(x)dx = 1.$$

Some continuous probability distributions			
Distribution	Density function	mean	Meaning or Relevance
Normal	$\rho(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}$	μ	It is an approximation to the sampling distribution of $\bar{X}$ for large n.
Student's t-distribution	$\rho(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$	$\mu = 0$	Distribution of the sample mean of n observations from a normal distribution relative to the true mean.
Chi-square	$\rho(x) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}\, x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$	$\mu = k$	Sum of the squares of independent normal standard variables.

The expected value and standard deviation of the distribution are:

$$\mu(X) = E(X) = \int_{-\infty}^{\infty} x\rho(x)dx, \qquad \sigma = \sqrt{\int_{-\infty}^{\infty} (x-\mu)^2 \rho(x)dx}.$$

The geometric interpretation of the probability $P(a < X < b)$: for a continuous random variable with distribution function ρ is the area under the curve $y = \rho(x)$ and above the x-axis when $a < x < b$. Notice that a or b or maybe both can be equal to ∞ or $-\infty$.

To find the probability that X fall in a given interval we use:

$$P(a < X < b) = \int_a^b \rho(x)dx.$$

9.1 The normal probability distribution and sampling distributions

The density function for the normal probability distribution satisfies the differential equation

$$\frac{dy}{dx} = -\frac{1}{\sigma^2}(x-\mu)y,$$

for some real numbers μ and σ ($\sigma > 0$). The solution to the equation is the family of functions

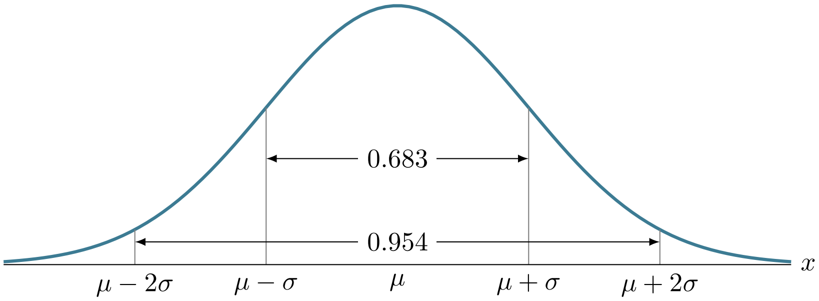
$$y = y(x) = Ke^{\frac{(x-\mu)^2}{2\sigma^2}},$$

and we are looking for the solution such that $\int_{-\infty}^{\infty} y(x)dx = 1$. Using the fact $\int_0^{\infty} e^{x^2/2} = \sqrt{2\pi}/2$ and the change of variable $x' = \frac{x-\mu}{\sigma}$, we get $K = \frac{1}{\sqrt{2\pi}\sigma}$ and the density function as

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

The expected value and standard deviation of the distribution are:

$$\mu = E(X) = \int_{-\infty}^{\infty} x\rho(x)dx, \qquad \sigma = \sqrt{E(X^2) - E(X)^2} = \sqrt{\int_{-\infty}^{\infty} (x-\mu)^2 \rho(x)dx}.$$



The maximum of the functions is at the point $(\mu, \frac{1}{\sqrt{2\pi}\sigma})$ and the inflexion points at $x = \mu \pm \sigma$. The normal curve with $\mu = 0$ and $\sigma = 1$ is called **standard normal distribution**. A random variable that follows a normal standard distribution is usually denoted with the letter Z and probabilities $P(a < Z < b)$ can be found in the **Standard Normal Distribution Table**.

The Standard Normal Distribution Table gives areas to the left, that is $P(Z < z)$. To find areas to the right and between two scores, you use:

$$P(Z > z) = 1 - P(Z < z) \quad \text{and} \quad P(z_1 < Z < z_2) = P(Z < z_2) - P(Z < z_1).$$

We have the raw score: $X = \sigma Z + \mu$.

The standard score: $Z = \dfrac{X - \mu}{\sigma}$.

Using a change a variable we can relate the raw and standard score proving that:

$$P(a < Z < b) = P(\frac{a-\mu}{\sigma} < X < \frac{b-\mu}{\sigma}).$$

A sampling distribution: is the probability distribution of a sample statistic based on all possible simple random samples of the same size from the population. For example the distribution of the sample mean $\bar{X}$ based on random samples of size n.

Whenever X is normally distributed with mean μ and s.dev.σ: the variable $\bar{X} = \dfrac{\sum x}{n}$ that averages random samples of size n is also normally distributed with mean and standard deviation given by the formulas:

$$\mu_{\bar{x}} = \mu_x, \qquad \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}.$$

Central limit Theorem: Regardless of the distribution followed by X with mean μ and standard deviation σ, the sequence of random variables $X_n = \bar{X}$ is close, for n large, to a normal distribution with mean μ and standard deviation $\dfrac{\sigma}{\sqrt{n}}$. For practical considerations $n \geq 30$ is usually sufficient.

As an application of the Central Limit Theorem to the sum of independent Bernoulli events: For $np > 5$ and $n(1-p) > 5$ we can use the continuous normal distribution with $\mu = np$ and $\sigma = \sqrt{np(1-p)}$ to approximate the binomial distribution with parameters n and p.

QUESTIONS FROM SECTION 9

1. Suppose that the random variable X follows a continuous probability distribution.
   (a) What is the probability $P(X = 1)$?

(b) If the probability $P(X > 3) = .3$, what is the probability $P(X < 3)$?

(c) What is the probability $P(-\infty < X < \infty)$?

2. Given the function $f(x)$, defined on the real numbers by the formulas:

$$f(x) = \begin{cases} 0 & x \leq 0 \\ x & 0 \leq x \leq 1 \\ 2 - x & 1 \leq x \leq 2 \\ 0 & 2 \leq x \end{cases}$$

(a) Show that $f(x)$ is the density function of a continuous probability distribution.

(b) Find the probability $P(-2 < X < 1)$.

(c) Find the probability $P(2 < X < 3)$.

3. Let z have the standard normal distribution. For each of the following probabilities, draw an appropriate diagram, shade the appropriate region and then determine the value:

(a) $P(0 < z < 1.74)$

(b) $P(0.62 < z < 2.48)$

(c) $P(z > 2.1)$

(d) $P(-1.31 < z < 1.07)$.

4. Let z have the standard normal distribution. For each of the following probabilities, draw an appropriate diagram, shade the appropriate region and then determine the value of z_c:

(a) $P(0 < z < z_c) = 0.4573$

(b) $P(z_c < z < 0) = 0.3790$

(c) $P(z < z_c) = 0.1190$

(d) $P(-z_c < z < z_c) = 0.8030$.

5. Let x be a normally distributed random variable with $\mu = 70$ and $\sigma = 8$. For each of the following probabilities, draw an appropriate diagram, shade the appropriate region and then determine the value:

(a) $P(70 < x < 80.4)$

(b) $P(61.2 < x < 85.2)$

(c) $P(x < 58)$

(d) $P(x > 76)$.

(e) $P(68 < \bar{x} < 72)$, if a random sample of size $n = 49$ is drawn.

(f) $P(\bar{x} > 71)$, if a random sample of size $n = 81$ is drawn.

6. Find z so that:

(a) 98% of the area under the standard normal curve lies between $-z$ and z.

(b) 97.5% of the area under the standard normal curve lies to the left of z.

(c) 46% of the area under the standard normal curve lies to the right of z.

7. Find the area under the standard normal curve

(a) between $z = -2.74$ and $z = 2.33$.

(b) between $z = -2.47$ and $z = 1.03$.

8. The lifetime of a certain type TV tube has a normal distribution with a mean of 80.0 and a standard deviation of 6.0 months. What portion of the tubes lasts between 62.0 and 95.0 months?

9. The scores in a standardized test are normally distributed with $\mu = 100$ and $\sigma = 15$.
   (a) Find the percentage of scores that will fall below 112.
   (b) A random sample of 10 tests is taken. What is the probability that their mean score $\bar{x}$ is below 112?

10. The weights (in pounds) of metal discarded in one week by households are normally distributed with a mean of 2.22 lb. and a standard deviation of 1.09 lb.
   (a) If one household is randomly selected, find the probability that it discards more than 2.00 lb. of metal in a week.
   (b) Find a weight p_{30} so that the weight of metal discarded by 70% of the houses is above x.

11. If the salary of computer technicians in the United States is normally distributed with the mean of \$32,550 and the standard deviation of \$2,000, find the probability for a randomly selected technician to earn
   (a) More than \$35,000.
   (b) Between \$31,500 and \$35,000.
   (c) What is the probability that the mean salary of a random sample of 4 technicians is more than \$35,000?

12. The lifetime of a AAA battery is normally distributed with mean $\mu = 28.5$ hours and standard deviation $\sigma = 5.3$ hours.
   (a) For a battery selected at random, what is the probability that the lifetime will be more than 30 hours.
   (b) For a sample of three batteries, what is the chance that all three last more than 30 hours?
   (c) For a sample of three batteries, what is the probability that their mean lifetime $\bar{x}$ is more than 30 hours?
   (d) What is the probability that the mean lifetime $\bar{x}$ of batteries from a package of 12 will be less than 27 hours?

13. In Jennifer's Fall 2014 history class, 14 of 34 students passed the class. If you assume a professor's passing rates are constant, would it be appropriate to use a normal curve approximation to the binomial distribution to estimate the mean passing rate for the same professor's Spring 2015 semester class of 28 students? Explain your answer.

14. According to the Vision Council of America, 75 percent of the U.S. adult population wears some form of glasses to correct their vision. In a random sample of 950 adults, what is the probability that fewer than 700 people wear glasses?

15. An environmental group did a study of recycling habits in a California community. It found that 70 percent of aluminum cans sold in the area were recycled. If 400 cans are sold in one day, what is the probability that between 260 and 300 will be recycled?

16. The weekly amount a family spends on groceries follows (approximately) a normal distribution with mean $\mu = \$200$ and a standard deviation $\sigma = \$15$.
   (a) If \$220 is budgeted for next week's groceries what is the probability that the actual cost will exceed the budget?
   (b) How much should be budgeted for weekly grocery shopping so that the probability that the budgeted amount will be exceeded is only 0.05?

10 Estimation: Confidence intervals and sample size

Estimation: is the process of inferring an unknown parameter using sample data. **A point estimation** for a parameter of the population is given by a single value of a statistic.

Given a population parameter θ and a sample statistic t representing a point estimate for θ, we will like to create an interval estimate with a high confidence of containing the actual parameter θ.
Let c be a real number $0 < c < 1$. The c-confidence interval for θ is an interval $[t - E_c, t + E_c]$ around t such that we will be $100c\%$ confident that it will cover the parameter θ of the entire population. **The statistic E_c is called the margin of error.** The critical value at level c for a continuous random variable X is a number x_c such that

$$P(-x_c < X < x_c) = c.$$

In other words $P(X < -x_c) = \frac{1-c}{2}$ or equivalently $P(X < x_c) = \frac{1+c}{2}$.

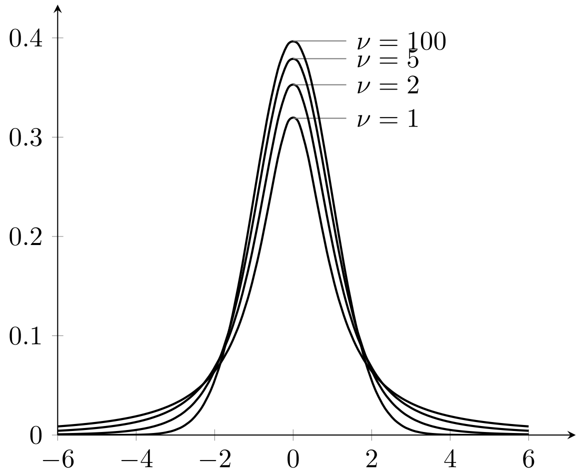
Confidence interval for the mean μ: In case we want to estimate the mean μ of the population using the statistic $\bar{x}$, the margin of error takes the shape:

Margin of Error when σ is known	$E_c = z_c \frac{\sigma}{\sqrt{n}}$
Margin of Error when σ is unknown	$E_c = t_c \frac{s}{\sqrt{n}}$

For samples of size n from a normal distribution of mean μ: the quotient of random variables

$$\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}},$$

follows a t-distribution. **The t-distribution depends on one parameter: the degrees of freedom ($d.f.$).** If we take a sample of n observations from a normal distribution, then the t-distribution with $d.f. = \nu = n - 1$ degrees of freedom can be defined as the distribution of the location of the sample mean relative to the true mean, divided by the sample standard deviation.



The density function is given by the formula with $d.f. = \nu$:

$$\rho(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

Where $\Gamma(x)$ is the Gamma function $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$ having the property that $\Gamma(n) = (n-1)!$

As the degrees of freedom grow larger and larger samples drawn from a normal population resemble more and more the whole population and the t-distribution get closer and closer to a normal standard distribution

The t-distribution is used to: estimate the μ of a **normal** distribution when the standard deviation σ is **unknown**.

Confidence interval for the proportion p: In case we want to estimate the proportion p of individuals on the population with a particular attribute using the sample proportion $\hat{p} = r/n$, as long as $n\hat{p} > 5$ and $n(1 - \hat{p}) > 5$, the margin of error can be computed with the formula:

$$E_c = Z_c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

QUESTIONS FROM SECTION 10

1. A study is being planned to estimate the mean number of semester hours taken by students at a college. The population standard deviation is assumed to be $\sigma = 4.7$ hours. How many students should be included in the sample to be 99% confident that the sample mean $\bar{x}$ is within one semester hour of the population mean μ for all students at this college?

2. To determine the mileage of a new model automobile, a random sample of 36 cars was tested. A sample with a mean of 32.6 mpg and a standard deviation of 1.6 mpg was obtained. Construct the 90% confidence interval for the actual mean mpg of the population of this model automobile.

3. A random sample of 12 employees was taken and the number of days each was absent for sickness was recorded (during a one-year period). If the sample had a mean $\bar{x}$ of 5.03 days and standard deviation s of 3.48 days, create a 95% confidence interval for the population mean days absent for sickness, assuming the distribution of absences is normal.

4. Computer Depot is a large store that sells and repairs computers. A random sample of 110 computer repair jobs took technicians an average of $\bar{x} = 93.2$ minutes per computer. Assume that σ is known to be 16.9 minutes. Find a 99% confidence interval for the population mean time μ for computer repairs.

5. The following data represent a sample of the number of home fires started by candles. Assuming that the number of home fires started by candles is approximately normally distributed find a 95% confidence interval for mean number of home fires started by candles each year.

   5400 5860 6070 6210 7360 8450 9960

6. Leonor decides to run for political office. In order for her name to appear on the ballot, she must collect 7,500 valid signatures from registered voters. After she collects 10,000 signatures, she decides to check what proportion of the ones she collected are valid. She takes a random sample of 150 of the signatures she collected and brings them to the Board of Elections to verify them. It turns out that of the sample of 150, only 87 are valid. Construct a 95 percent confidence interval for the proportion of valid signatures she has collected.

7. In a Gallup poll, 1025 randomly selected adults were surveyed. 400 of them said that they shopped on the internet at least a few times per year. Construct a 99 percent confidence interval to estimate the percentage of all adults who shop on the internet several times per year.

8. A random sample of 41 NBA players gave a standard deviation $s = 3.32$ inches for their height. How many more NBA players have to be included in the sample to make 95% sure that the sample mean $\bar{x}$ of their height is within 0.75 inch of the mean μ of the height of the population of all NBA players.

9. A 99% confidence interval for the mean number μ of televisions per American household is $(.92, 4.97)$. For each of the following statements about the above confidence interval, choose true or false and explain your answer:

   (a) The probability that μ is between .92 and 4.97 is .99.

   (b) We are 99% confident that the true mean number μ of televisions per American household is between .92 and 4.97

   (c) 99% of all samples should have $\bar{x}$ between .92 and 4.97.

   (d) 99% of all American households have between .92 and 4.97 televisions.

   (e) Of many intervals calculated the same way (99% intervals), we expect 99% of them to capture the population mean μ.

   (f) Of many intervals calculated the same way (99% intervals), we expect 100% of them to capture the sample mean $\bar{x}$.

10. A study of 40 English composition professors showed that they spent, on average, 12.6 minutes correcting a student's term paper. Find the 90% confidence interval of the mean time for all composition papers when $\sigma = 2.5$ minutes. If we change to do the 95% confidence interval instead of the 90%, without doing the calculations, do you expect the new interval to be bigger or smaller than before? Explain your answer.

11 Testing statistical hypothesis

Hypothesis testing: is a statistical test to decide whether or not there is enough evidence in a sample data to infer that some conclusion is true for the whole population.

H_0: The null hypothesis. The statement under investigation, that is usually a statement of "no effect" or "no difference". It represents a statement that we expect to reject.

H_1: The alternate hypothesis. An alternate to the null hypothesis that we expect to adopt if the evidence is enough to reject H_0.

The p-value: is the probability that we observe results as extreme as the test statistic observed if the null hypothesis H_0 were to be true.

Error of type I: Is the probability α that we reject H_0 when it was in fact true. It represents our willingness of rejecting a true null hypothesis. The number α is also called **the significance level** of the test. An outcome will be considered "unlikely" if its probability is less than α.

Error of type II: Is the probability β of accepting H_0 when it was in fact false.

The probability of rejecting H_0 when it was in fact false: is the quantity $1 - \beta$ and is call the power of the test.

By increasing the significance level α: we are more likely to reject the null hypothesis. This means that we are less likely to accept the null hypothesis when it is false; i.e., less likely to make a Type II error. Hence, the power of the test is increased.

Sample Test Statistics of Test of Hypothesis:

Test Statistic for μ (σ known): $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

Test Statistics for μ (σ unknown): $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$

The rejection zone: Is the portion of the x-axis that represents values as extreme as the level of significance α. If test statistic falls in the rejection zone it means that the probability of observing such an extreme result when H_0 is correct is less than α (p-value $< \alpha$) and we conclude that H_0 should be rejected. Otherwise if the test statistics does not fall in the rejection zone or critical zone (equivalently p-value $> \alpha$), we conclude that there are not enough evidence to reject H_0.

QUESTIONS FROM SECTION 11

1. Gregor Mendel was a pioneer in the theory of genetics. His idea was to assign probabilities to significant population traits of plants or animals, like eye color, based on "dominant" or "recessive" traits. For example, he studied peas with green pods (a dominant trait) or yellow pods (a recessive trait). He predicted that the probability that a hybrid ("offspring") of a green pea with a yellow pea will have a yellow pod is $p = 0.25$.

   Mendel conducted an experiment of green-yellow hybrids. In one experiment, 428 offspring had green pods and 152 offspring had yellow pods.

   Use a level of significance of $\alpha = 0.01$ to test the claim that Mendel's claim that $p = 0.25$ is wrong.

2. A teacher has developed a new technique for teaching which he wishes to check by statistical methods. If the mean of a class test turns out to be 60 (or less), the results will be considered unsuccessful. Alternatively, if the mean is greater than 60, the results will be considered successful. The results of the test with a class of 36 students had a mean $\bar{x} = 66.2$ with a standard deviation of $s = 24.0$. Test whether the results were successful at the $\alpha = 5\%$ level of significance. (Use 1-tail test.) State the null and the alternate hypothesis and include diagrams.

3. The average annual salary of employees at a retail store was \$28,750 last year. This year the company opened another store. Suppose a random sample of 18 employees had an average annual salary of $\bar{x} = \$25,810$ with sample standard deviation of $s = \$4230$. Use a level of significance $\alpha = 1\%$ to test the claim that the average annual salary for all employees is different from last years average salary. Assume salaries are normally distributed.

4. A machine in the lodge at a ski resort dispenses a hot chocolate drink. The average cup of hot chocolate is supposed to contain $\mu = 7.75$ ounces. We may assume that x has a normal distribution with $\sigma = 0.3$ ounces. A random sample of 16 cups of hot chocolate from this machine had a mean content of $\bar{x} = 7.62$ ounces. Use a $\alpha = 0.05$ level of significance and test whether the mean amount of liquid is different than 7.75 ounces.

5. A teacher has developed a new technique for teaching which she wishes to check by statistical methods. If the mean of a class test turns out to be 70 (or less), the results will be considered unsuccessful. Alternatively, if the mean is greater than 70, the results will be considered successful. State the null and the alternate hypothesis (Use 1-tail test).

6. A Type II error is made when

   (a) the null hypothesis is accepted when it is false.

   (b) the null hypothesis is rejected when it is true.

(c) the alternate hypothesis is accepted when it is false.

(d) the null hypothesis is accepted when it is true.

(e) the alternate hypothesis is accepted when it is true.

7. A Type I error is made when

   (a) the null hypothesis is accepted when it is false.

   (b) the null hypothesis is rejected when it is true.

   (c) the alternate hypothesis is accepted when it is false.

   (d) the null hypothesis is accepted when it is true.

   (e) the alternate hypothesis is accepted when it is true.

8. What is the effect of increasing the sample size in the type I and type II errors?

9. How many Kleenex should a package of tissues contain? Researchers determined that 60 tissues is the average number of tissues used during a cold. Suppose a random sample of 100 Kleenex users yielded the following data on the number of tissues used during a cold: $\bar{x} = 52$, $s = 22$. Using the sample information provided, calculate the value of the test statistic t.

10. A pharmaceutical company claims that its weight loss drug allows women to lose in average of $\mu = 8lb$ after one month of treatment. If we want to conduct an experiment to determine if the patients are losing less weight than advertised, what would be the null H_0 and alternative hypothesis H_1?

11. Suppose our p-value is .047. What will our conclusion be at alpha levels of $\alpha = .10$, $\alpha = .05$ and $\alpha = .01$? Explain your selection.

    (a) We will reject H_0 at $\alpha = .10$, but not at $\alpha = .05$

    (b) We will reject H_0 at $\alpha = .10$ or .05, but not at $\alpha = .01$

    (c) We will reject H_0 at $\alpha = .10$, .05, or .01

    (d) We will not reject H_0 at $\alpha = .10$, .05, or .01

12. Suppose the p-value for a test is .02. Which of the following is true? Explain your selection.

    (a) We will not reject H_0 at $\alpha = .05$

    (b) We will reject H_0 at $\alpha = .01$

    (c) We will reject H_0 at $\alpha = 0.05$

    (d) We will reject H_0 at alpha equals 0.01, 0.05, and 0.10

    (e) None of the above is true.

13. A survey was conducted to get an estimate of the proportion of smokers among the graduate students. Report says 35% of them are smokers. Lida doubts the result and thinks that the actual proportion is much less than this. Choose the correct choice of null and alternative hypothesis Lida wants to test. Explain your selection.

    (a) $H_0 \colon p = .35$ versus $H_1 : p \neq .35$.

    (b) $H_0 \colon p = .35$ versus $H_1 : p > .35$.

    (c) $H_0 \colon p = .35$ versus $H_1 : p < .35$.

    (d) None of the above

14. The null hypothesis $H_0 : \mu = .5$ against the alternative $H_1 : \mu > .5$ was rejected at level $\alpha = 0.01$. Pete wants to know what the test will result at level $\alpha = 0.10$. What will be his conclusion? Explain your selection.

    (a) Reject H_0.

(b) Fail to Reject H_0.

(c) No conclusion can be made.

(d) Reject H_1.

15. The null hypothesis $H_0 : \mu = 5$ against the alternative $H_1 : \mu > 5$ was rejected at certain level of significance. What will be the conclusion for testing $H_0 : \mu = 5$ against the alternative $H_1 : \mu \neq 5$ at the same level? Explain your selection.

    (a) Fail to Reject H_0.

    (b) Reject H_0.

    (c) No conclusion can be made.

    (d) Reject H_1.

16. A researcher wanted to test the null hypothesis $H_0 : \mu = 10$ vs. $H_1 : \mu > 10$. She obtained that a sample statistic $\bar{x} = 10.5$ with a sample size of $n = 20$ did not provide enough evidence to reject H_0 at a significance level $\alpha = .01$. What can we say about the conditional probability

$$p = Pr(\bar{x} \geq 10.5 \mid \mu = 10)?$$

    Explain your answer.

    (a) $p < .01$

    (b) $p > .01$

    (c) Both (a) and (b) can occur.

    (d) $p = .01$

17. A null hypothesis was rejected at level$\alpha = 0.10$. What will be the result of the test at level $\alpha = 0.05$? Explain your answer.

    (a) Reject H_0.

    (b) Fail to Reject H_0.

    (c) No conclusion can be made.

    (d) Reject H_1.

12 Inferences about differences

Two samples are dependent: if each data value in one sample can be paired with a corresponding value of the other sample.

Consider a random sample of n pairs: assuming that the differences d between the first and second members of each pair follow an approximately normal distribution (this would be true for n big), then the random variable

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}},$$

will follow a t-distribution with $n - 1$ degrees of freedom.

Two samples are independent: if the selection of sample data from one population is completely unrelated to the selection from the other population.

Differences of the means when σ_1, σ_2 are known values: Suppose that x_1 and x_2 are normally distributed with means μ_1 and μ_2 and standard deviations σ_1 and σ_2. If we take independent random samples of size n_1 and n_2 respectively from the x_1 and the x_2 distributions, the variable $\bar{x}_1 - \bar{x}_2$ will follow a normal distribution with mean

$$\mu = \mu_1 - \mu_2, \text{ and standard deviation } \sigma = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Confidence interval for $\mu_1 - \mu_2$: $(\bar{x}_1 - \bar{x}_2) - E < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + E,$

where E is given by: $E = z_c\sqrt{\dfrac{\sigma^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}.$

Differences of the means when σ_1, σ_2 are unknown: Suppose that x_1 and x_2 are normally distributed with means μ_1 and μ_2. If we take independent random samples of size n_1 and n_2 respectively from the x_1 and the x_2 distributions obtaining sample standard deviations s_1 and s_2 for our samples, the sample test statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_1 - \mu_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

will follow approximately a Student's t distribution with degrees of freedom equals $\min(n_1 - 1, n_2 - 1)$. A more accurate value for the degrees of freedom can be found using Satterthwaite's approximation:

$$d.f. \approx \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{1}{n_1 - 1}\left(\dfrac{s_1^2}{n_1}\right)^2 + \dfrac{1}{n_2 - 1}\left(\dfrac{s_2^2}{n_2}\right)^2}.$$

QUESTIONS FROM SECTION 12

1. For a random sample of 36 data pairs, the sample mean of the differences is 0.8 and the standard deviation of the differences is 2. Test the claim that the population mean of the differences is different from 0 at a 5% level of significance.

2. We have a data consisting of $n = 9$ pairs of observation with sample mean and standard deviation of $\bar{d} = 33.3$ and $s = 22.9$. At the level of significance $\alpha = .01$, test the claim that the mean of the differences is positive.

3. Two samples of size $n_1 = 10$ and $n_2 = 12$ from two normally distributed populations of unknowns means μ_1 and μ_2 gave sample statistics $\bar{x}_1 = 11$ and $\bar{x}_2 = 10$. Assume that the standard deviations are known though to be $\sigma_1 = 2.5$ and $\sigma_2 = 3$. Can we say that there are significative difference between the means of the populations with $\alpha = .05$? Build the 95% confidence interval for $\mu_1 - \mu_2$.

4. The result of experimenting with two normally distributed populations gives:

Sample from x_1	$\bar{x}_1 = 20$	$s_1 = 8.5$	$n_1 = 13$
Sample from x_2	$\bar{x}_2 = 11$	$s_2 = 7.5$	$n_2 = 10$

   (a) Test at the 5% significance level the claim that $m_2 > m_1$.

   (b) Find the 95% confidence interval for $\mu_1 - \mu_2$.

   (c) Compare the results in (a) and (b).

13 The chi-square distribution

The chi-square distribution $\chi^2(k)$:, with $d.f. = k$ degrees of freedom is the distribution that follows the sum of the squares of k independent standard normal random variables.

In mathematical terms: if $z_1, z_2, \ldots, z_k$ are independent standard normal variables, the sum of their squares

$$\sum_{i=1}^{k} z_i^2 \quad \sim \quad \chi^2(k).$$

The $\chi^2(k)$ distribution: has positive real numbers as domain and it is not symmetrical. The mean is always k and as long as $k > 2$, the mode is always at $k - 2$.

The chi-squared distribution is used primarily for: χ^2 test of independence in contingency tables and the χ^2 test of goodness of fit of observed data to hypothetical distributions.

In a test of independency: in a contingency table, the **number of observations of type** i is denoted by O_i. **the expected frequency** E_i of type i is given by

$$E_i = \frac{(\text{Row Total for } i)(\text{Column Total for } i)}{\text{Sample Size}},$$

and **the statistic**

$$\sum_{i=1}^{n} \frac{O_i - E_i}{E_i} \quad \sim \quad \chi^2((R-1)(C-1)),$$

where R and C represents the number of rows and columns in our table.

The χ^2 goodness of the fit test: determines how well a theoretical distribution (such as normal, binomial, Poisson or simply a prescribed distribution) fits an empirical distribution. In the goodness of the fit test, the population is divided in categories and a theoretical probability or frequency is assigned to each category. Then we get a random sample of size n and count the amount of observed values n_i in each category.

The observed frequency of the category i: is denoted by $O_i = n_i/n$ and **the expected frequency** by $E_i = np_i$, where n is the sample size and p_i is theoretical probability of the category i.

The statistic:

$$\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad \sim \quad \chi^2(n-1)$$

The Poisson distribution: describes the number of successes in blocks of time or space, when it is assumed that successes happen independently of each other and with equal probability at each point.

When X follows a Poisson with mean μ: the probability of X is successes is determined by the formula:

$$P(X \text{ successes}) = \frac{e^{-\mu}\mu^X}{X!},$$

where μ is the mean number of independent successes in a unit of time or space.

The Poisson distribution: is a useful tool to determine whether events or objects occur randomly in space or time. When events are random in time or space (not clumped or disperse) it is reasonable to think that they will follow a Poisson distribution.

1. Test the independency of the factors A, B with the factors γ, β and δ and the .01 level of significance.

	A	B	Row Total
γ	62	45	107
β	68	94	162
δ	56	81	137
Column Total	186	220	406

2. The table bellow represents represents the number of boys in families of two children. Assume that the sex of consecutive sons is independent. Test the hypothesis of the number of sons following a binomial distribution with mean $n = 2$ and $p = .51$. In case we do not have p our best guess will be the value of the estimate $\hat{p}$. Use $\alpha = .05$.

Number of boys	Number of families
0	217
1	545
2	238
Total	1000

3. Test the claim that the numbers in the table with the given frequencies follow a Poisson distribution with mean $\mu = 2.44$. (this is equivalent to test for the randomness of the numbers and frequencies in the table) Use $\alpha = .05$.

Number	Frequency
0	7
1	6
2	2
3	3
4	4
5	6
6	1
Total	29

14 ANSWERS

14.1 Answers to problems in section 1

1. A. Nominal. B. Ratio. C. Ratio. D. Ordinal. E. Ratio. F. Interval.

14.2 Answers to problems in section 2

1. The class width has to be 6. We then have the following frequency table.

Class Limits	Class Boundaries		Class Marks
Lower-Upper	Lower-Upper	Frequency	(midpoints)
30 – 35	29.5 – 35.5	3	32.5
36 – 41	35.5 – 41.5	7	38.5
42 – 47	41.5 – 47.5	11	44.5
48 – 53	47.5 – 53.5	4	50.5

And we have the following histogram: Figure 1

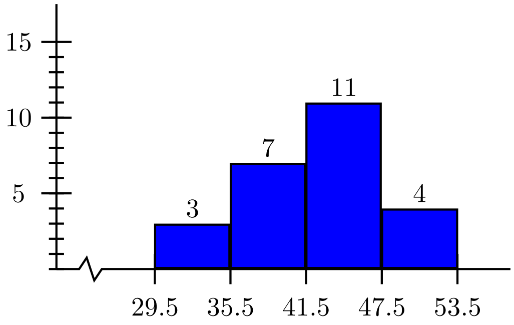


Figure 1: The histogram of problem 1 in section 2

14.3 Answers to problems in section 3

1. The range is 12, the mode is 56, the mean is $\mu = 53$, the standard variation is $\sigma = 3.69$, the variance is $\sigma^2 = 13.6$. The quartiles are $Q_1 = 50$, the median $Q_2 = 52.5$, and $Q_3 = 56$ while the interquartile range is 6.
2. Mean is $\bar{x} = 9.5$, range is 9, sample standard deviation is $s = 2.64$.
3. The range is 32.9. The median is 22. The mean is 26. The standard deviation is $s = 11.22$.
4. A. Mean. B. Median. C. Mode.
5. 4 students. No, she cannot.
6. Approximately 25% of the data.

14.4 Answers to problems in section 4

1. $[77, 93]$.
2. A. $[0, 88]$ B. $[0, 113.6]$ C. $[0, 138.4]$.
3. $k = \sqrt{2}$.
4. $k = \sqrt{3}$.

14.5 Answers to problems in section 5

1. The correlation coefficient is $r = 0.9$. The line of least squares is $\hat{y} = 0.3 + 0.09x$. For a region with disposable annual income of \$25,000,000 the model predicts sale of 2,550 cases. The scatter graph and the plot of the line are shown in Figure 2.
2. A. (d) B. (b) C. (a) D. (c).

14.6 Answers to problems in section 6

1. A. 0.175 B. 0.875 C. 0.5.
2. A. $\frac{1}{36}$ B. $\frac{1}{36}$ C. $\frac{1}{18}$ D. $\frac{1}{12}$ E. $\frac{1}{9}$ F. $\frac{11}{36}$ G. $\frac{1}{4}$ H. $\frac{1}{2}$.
3. A. 120 B. 220.
4. A. $\frac{5}{33}$ B. $\frac{1}{11}$ C. $\frac{2}{11}$.

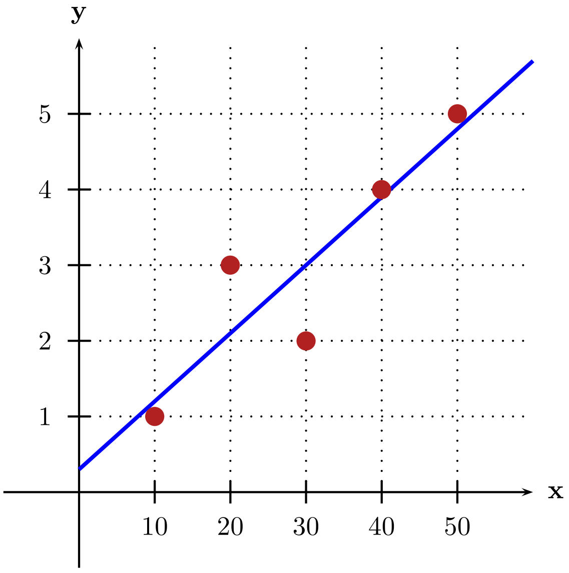


Figure 2: The scatter plot and the regression line of problem 1 in section 5

5. A. $\frac{25}{144}$ B. $\frac{1}{12}$ C. $\frac{1}{6}$.
6. A. $\frac{2}{17}$ B. $\frac{13}{34}$ C. $\frac{11}{850}$ D. $\frac{169}{1700}$ E. $\frac{1}{5525}$ F. $\frac{6}{5525}$.
7. A. $\frac{1}{20}$ B. $\frac{1}{8}$ C. $\frac{25}{28}$ D. $\frac{25}{27}$ E. $\frac{3}{8}$.
8. A. $\frac{421}{1259}$ B. $\frac{1112}{1259}$ C. $\frac{96}{1259}$ D. $\frac{608}{1259}$ E. $\frac{430}{1259}$ F. $\frac{79}{321}$ G. $\frac{79}{300}$.
9. A. $\frac{145}{208}$ B. $\frac{63}{208}$.
10. A. 0.09 B. 0.21.

14.7 Answers to problems in section 7

1. The expected value of the distribution is $\mu = 3.9$ and the standard deviation is $\sigma = 1.37$. The graph of the distribution is Figure 3

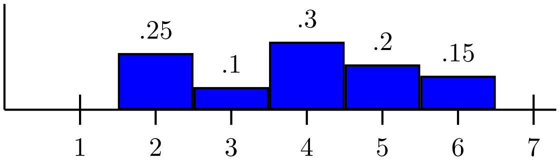


Figure 3: The graph of the probability distribution of problem 1 in section 7

2. The expected value is $\mu = 3.05$ and the standard deviation $\sigma = 1.58$.
3. First compute the probabilities (you can also get these values from the tables in the appendix of the textbook):

x	0	1	2	3	4	5	6	7
$P(x)$	.008	.055	.164	.273	.273	.164	.055	.008

The graph is Figure 4:

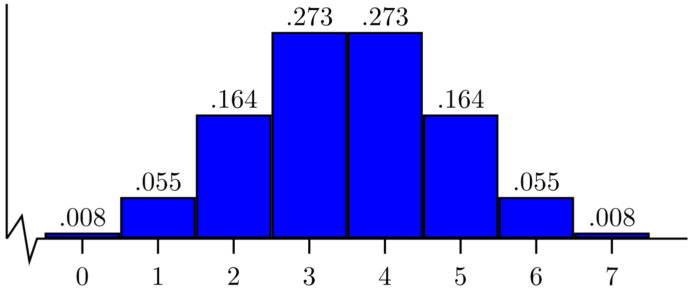


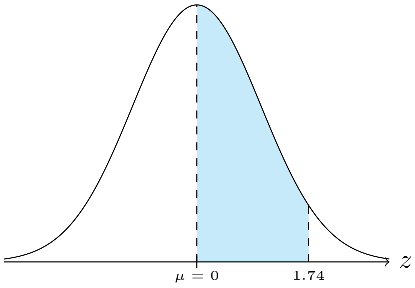
Figure 4: The graph of the binomial distribution of problem 3 in section 7

14.8 Answers to problems in section 8

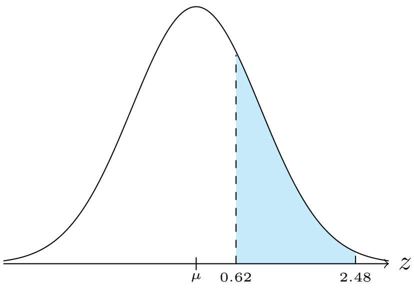
1. $P(0 \le r \le 4) \approx 0.63$.
2. A. $P(0 \le r \le 7) = 0.951$ B. $P(r = 7) = .081$ C. $P(5 \le r \le 15) = 0.485$ D. $P(0 \le r \le 2) = 0.128$ E. $P(10 \le r \le 15) = 0.004$.
3. A. $P(7 < r \le 15) = 0.951$. Why is this answer the same as the answer for 26 (a)? B. $P(7 < r \le 10) = 0.382$.
4. A. $P(0 \le r \le 1) = 0.881$ B. $P(0 \le r < 3) = 0.98$ C. The expected number $\mu = 0.6$ and the standard deviation $\sigma = 0.755$.
5. The expected number $\mu = 4.5$ and the standard deviation $\sigma = 1.061$.
6. $P(r = 0) = \dfrac{8}{27}$.

14.9 Answers to problems in section 9

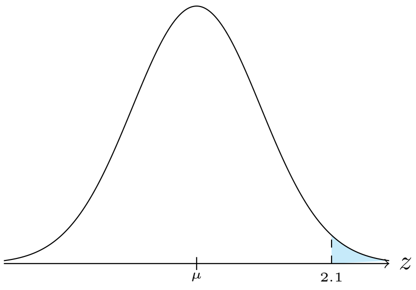
1. (a) $P(X = 1) = 0$
   (b) $P(X < 3) = .7$
   (c) $P(-\infty < X < \infty) = 1$.
2. (a) Total area is $A = 2(1)/2 = 1$.
   (b) $P(-2 < X < 1) = .5$.
   (c) $P(2 < X < 3) = 0$.
3. A. $P = 0.4591$



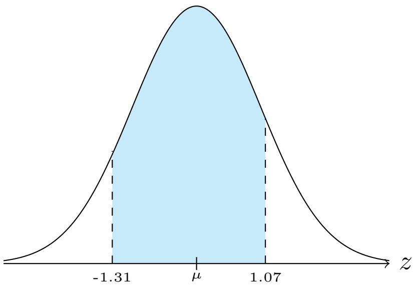
B. $P = 0.2610$



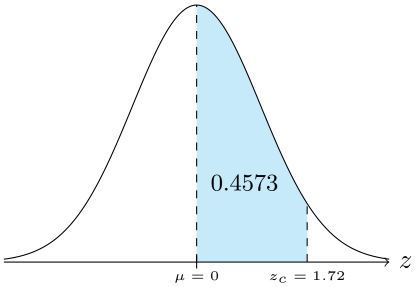
C. $P = 0.0179$



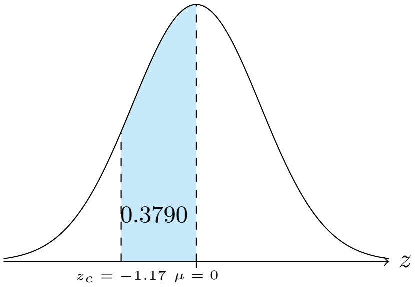
D. $P = 0.7626$



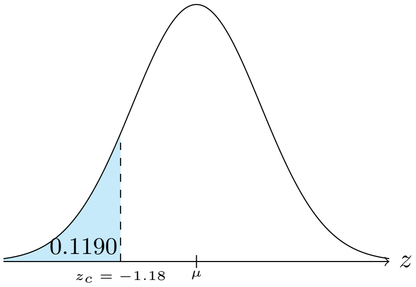
4. A. $z_c = 1.72$



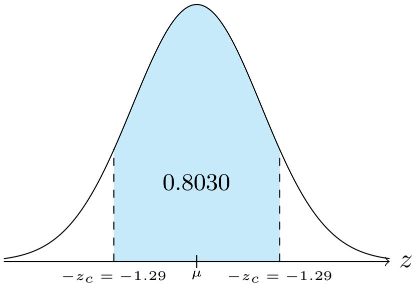
B. $z_c = -1.17$



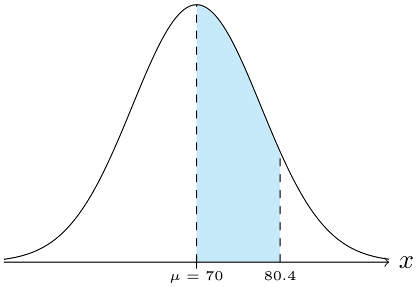
C. $z_c = -1.18$



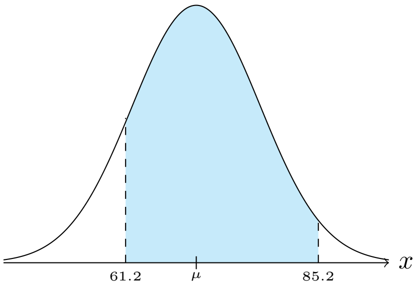
D. $z_c = 1.29$



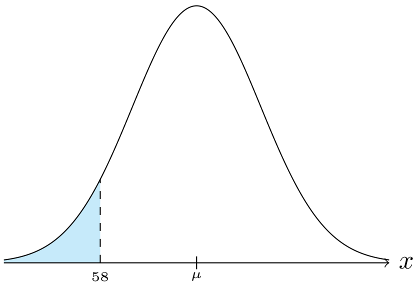
5. A. $P = 0.4032$



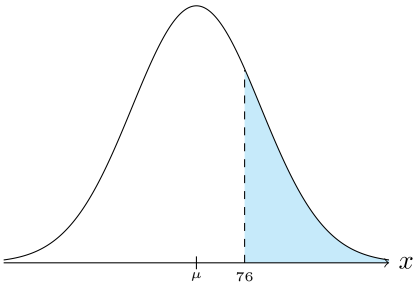
B. $P = 0.8356$



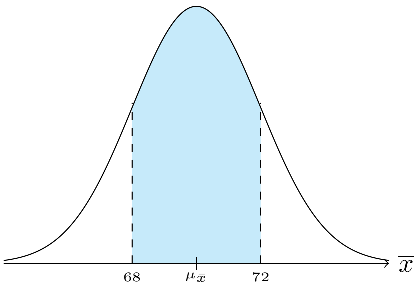
C. $P = 0.0668$



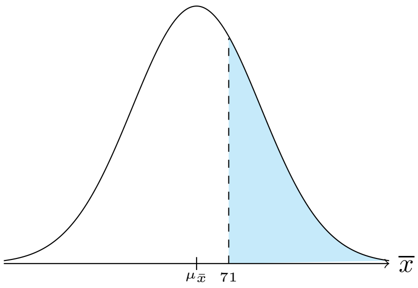
D. $P = 0.2266$



E. $P = 0.9198$



F. $P = 0.1292$



6. A. $z = 2.33$ B. $z = 1.96$ C. $z = 0.1$.

7. A. 0.987 B. 0.8417.

8. 99.24%.

9. A. 78.81% B. 0.9943.

10. A. $P(x > 2.00) = 0.58$ B. 1.65 lb.

11. A. 0.1093 B. 0.5926 C. 0.0071.

12. A. 0.3897 B. 0.0592 C. 0.3121 D. 0.1635.

13. A binomial distribution can be approximated by a normal distribution if both $np > 5$ and $nq > 5$. In Fall 2014 the passing rate was $p = 0.41$ with $np = 14 > 5$ and $nq = 20 > 5$ so it would be appropriate to assume a normal distribution for the next semester as well. As in Spring 2015 $n = 28$, using the normal distribution approximation would be assuming that $0.18 \leq p \leq 0.82$; $p = 0.41$ is within this range.

14. 0.1660

15. 0.9750

16. A. 0.0918 B. \$224.67.

14.10 Answers to problems in section 10

1. 148.

2. $[32.16, 33.04]$.

3. $[2.82, 7.24]$.

4. $[89.04, 97.36]$.

5. $[5518.54, 8570.04]$.

6. $0.50 < p < 0.66$

7. $0.35 < p < 0.43$

8. 35 more players need to be included.

14.11 Answers to problems in section 11

1. **Partial solution:** $H_0 : p = 0.25$, $H_\mathrm{a} : p \neq 0.25$. $n = 428 + 152$ so $\hat{p} = 0.26$. The sample test statistic is

$$z = \frac{0.26 - 0.25}{\sqrt{\dfrac{(0.25)(0.75)}{580}}} = \frac{0.01}{0.01798} = 0.56$$

$$P\text{-value } = 2 \cdot P(z \leq -0.56) = 0.5754 > 0.01 = \alpha$$

   Conclusion: Do not reject H_0. The results were not statistically significant at the 1% level of significance. Based on the sample data, we think that the probability that a pea hybrid will have a yellow pod is 0.25.

2. **Partial solution:** $H_0 : \mu = 60$ (or $\mu \leq 60$), $H_\mathrm{a} : \mu > 60$. The critical z-value is $z_c = 1.645$. Then

$$z = \frac{66.2 - 60.0}{\dfrac{24.0}{\sqrt{36}}} = \frac{6.2}{4.0} = 1.55 < z_c$$

   Conclusion: Do not reject H_0. The results were statistically unsuccessful at the 5% level of significance. (That is the results could not be distinguished from a random sample from a normal population with mean $\mu = 60$ and standard deviation $\sigma = 24.0$.)

3. **Partial solution:** $H_0 : \mu = 28,750$, $H_\mathrm{a} : \mu \neq 28,750$. The test statistic is

$$t = \frac{25810 - 28750}{\dfrac{4230}{\sqrt{18}}} = -\frac{2940}{997.02} = -2.949.$$

   For $d.f. = 17$, the test statistic $t = -2.949$ is in the interval

$$2.898 < |t| < 3.965.$$

   Thus a 2-tail test shows

$$0.010 > P\text{-value} > 0.001.$$

   Conclusion: Reject H_0 because the P-value $< \alpha = 0.01$. At the 1% level of significance, the evidence is sufficient to reject H_0. Based on the sample data, we think that the mean annual salary is different from that of the previous year.

4. **Partial solution:** $H_0 : \mu = 7.75$, $H_a : \mu \neq 7.75$. Critical value $\pm z_c = \pm 1.96$. Then

$$z = \frac{7.62 - 7.75}{\frac{0.3}{\sqrt{16}}} = -1.73 > -z_0.$$

   Conclusion: Not enough information to reject H_0.

5. $H_0 \colon \mu = 70$
   $H_1 \colon \mu > 70$.

6. (a)

7. (b)

8. The probability of making an error of type I will not be affected, while the probability of an error of type II will decrease (the power of the test increases).

9. $t = -3.64$

10. $H_0 \colon \mu = 8; H_1 \colon \mu < 8$

11. (b) 12.(c) 13.(c) 14.(a) 15.(c) 16.(b) 17.(b)