BarCharts, Inc.®
WORLD'S #1 ACADEMIC OUTLINE
Quick Study® ACADEMIC

Statistics EQUATIONS & ANSWERS™

Essential Tools for Understanding Statistics & Probability – Rules, Concepts, Variables, Equations, HARD & EASY Problems, Helpful Hints & Common Pitfalls

DESCRIPTIVE STATISTICS

Methods used to simply describe data set that has been observed

KEY TERMS & SYMBOLS

quantitative data: data variables that represent some **numeric quantity** (is a numeric measurement).

categorical (qualitative) data: data variables with values that reflect some **quality** of the element; one of several categories, not a numeric measurement.

population: "the **whole**"; the entire group of which we wish to speak or that we intend to measure.

sample: "the **part**"; a representative subset of the population.

simple random sampling: the most commonly assumed method for selecting a sample; samples are chosen so that every possible sample of the same size is equally likely to be the one that is selected.

N: size of a population.

n: size of a sample.

x: the value of an observation.

f: the frequency of an observation (i.e., the number of times it occurs).

frequency table: a table that lists the values observed in a data set along with the frequency with which it occurs.

(population) parameter: some numeric measurement that describes a population; generally not known, but **estimated** from sample statistics.
EX: *population mean:* **μ**; *population standard deviation:* **σ**; *population proportion:* **p** (sometimes denoted π)

(sample) statistic: some numeric measurement used to describe data in a sample, used to estimate or make inferences about population parameters.
EX: *sample mean:* $\bar{x}$; *sample standard deviation:* **s**; *sample proportion:* $\hat{p}$

Sample Problems & Solutions

1. A student receives the following exam grades in a course: 67, 88, 75, 82, 78

a. Compute the mean: $\bar{x} = \frac{\sum x}{n} = \frac{67+88+75+82+78}{5} = \frac{390}{5} = \mathbf{78}$

b. What is the median exam score?
in order, the scores are: 67, 75, 78, 82, 88; middle element = **78**

c. What is the range? range = maximum – minimum = 88 – 67 = **21**

d. Compute the standard deviation:

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{(67-78)^2+(88-78)^2+(75-78)^2+(82-78)^2+(78-78)^2}{4}} = \sqrt{\frac{246}{4}} = \sqrt{61.5} = \mathbf{7.84}$$

e. What is the *z* score for the exam grade of 88? $z = \frac{x-\bar{x}}{s} = \frac{88-78}{7.84} = \frac{10}{7.84} = \mathbf{1.28}$

2. The residents of a retirement community are surveyed as to how many times they've been married; the results are given in the following frequency table:

						Sums
x = # of marriages	0	1	2	3	4	n/a
f = # of observations	13	42	37	12	6	$110 = n$
xf	0	42	74	36	24	176

a. Compute the mean: $\bar{x} = \frac{\sum xf}{n} = \frac{176}{110} = \mathbf{1.6}$

b. Compute the median: Since $n = \sum f = 110$, an even number, the median is the average of the observations with ranks $\frac{n}{2}$ and $\frac{n}{2}+1$ (i.e., the 55th and 56th observations)

While we could count from either side of the distribution (from 0 or from 4), it is easier here to count from the bottom: The first 13 observations in rank order are all 0; the next 42 (the 14th through the 55th) are all 1; the 56th through the 92nd are all 2; since the 55th is a 1 and the 56th is a 2, the median is the average: (1 + 2) / 2 = **1.5**

c. Compute the IQR: To find the IQR, we must first compute Q1 and Q3; if we divide *n* in half, we have a lower 55 and an upper 55 observations; the "median" of each would have rank $\frac{n+1}{2}$ = 28; the 28th observation in the lower half is a 1, so Q1 = 1 and the 28th observation in the upper half is a 2, so Q2 = 2; therefore, IQR = Q3 – Q1 = 2 – 1 = **1**

Formulating Hypotheses

Type	Statistic	Formula	Important Properties
measures of center (measures of central tendency) *indicate which value is typical for the data set*	**mean**	from raw data $\bar{x} = \frac{\sum x}{n}$; from a frequency table $\bar{x} = \frac{\sum xf}{n}$	sensitive to extreme values; any outlier will influence the mean; **more useful for symmetric data**
	median *the middle element in order of rank*	*n odd:* median has rank $\frac{n+1}{2}$; *n even:* median is the average of values with ranks $\frac{n}{2}$ *and* $\frac{n}{2}+1$	not sensitive to extreme values; **more useful when data are skewed**
	mode	the observation with the highest frequency	only measure of center **appropriate for categorical data**
	mid-range	$\frac{maximum+minimum}{2}$	not often used; highly sensitive to unusual values; **easy to compute**
measures of variation (measures of dispersion) *reflect the variability of the data (i.e., how different the values are from each other)*	**sample variance**	$s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$	not often used; units are the **squares** of those for the data
	sample standard deviation	$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$	square root of variance; **sensitive to extreme values;** commonly used
	interquartile range (IQR)	IQR = Q3 – Q1 (*see* **quartile,** below)	less sensitive to extreme values
	range	maximum – minimum	not often used; **highly sensitive to unusual values;** easy to compute
measures of relative standing (measures of relative position) *indicate how a particular value compares to the others in the same data set*	**percentile**	data divided into 100 equal parts by rank (i.e., the k^{th} percentile is that value **greater than** *k*% of the others)	important to apply to ***normal*** **distributions** (*see* **probability distributions**)
	quartile	data divided into 4 equal parts by rank: Q3 (third quartile) is the value greater than ¾ of the others; Q1 (first quartile) is greater than ¼; Q2 is identical to the median	used to compute IQR (*see* **IQR,** above); Q3 is often viewed as the "median" of the upper half, and Q1 as the "median" of the lower half; Q2 is the median of the data set
	z score	$z = \frac{x-\bar{x}}{s}$ to find the value of some observation, *x*, when the *z* score is known: $x = \bar{x} + zs$	measures the distance from the mean in terms of standard deviation

PROBABILITY

Examples of Sample Spaces

Probability Experiment	Sample Space
toss a fair coin	{heads, tails} or {H, T}
toss a fair coin twice	{HH, HT, TH, TT} *there are* ***two*** *ways to get heads just once*
roll a fair die	{1, 2, 3, 4, 5, 6}
roll two fair dice	{(1,1), (1,2), (1,3). . . (2,1), (2,2), (2,3). . . (6,4), (6,5), (6,6)} *a total of* ***36 outcomes:*** six for the first die, times another six for the second die
have a baby	{boy, girl} or {B, G}
pick an orange from one of the trees in a grove, and weigh it	{some positive real number, in some unit of weight} *this would be a* ***continuous*** *sample space*

KEY TERMS & SYMBOLS

probability experiment: any process with an outcome regarded as random.

sample space (S): the set of all possible outcomes from a probability experiment.

events (A, B, C, etc.): subsets of the sample space; *many problems are best solved by a careful consideration of the defined events.*

P(A): the probability of event A; ***for any event A,*** **$0 \le P(A) \le 1$,** and for the entire sample space S, **P(S) = 1**

"equally likely outcomes": a very common assumption in solving problems in probability; if all outcomes in the sample space S are equally likely, then the probability of some event A can be calculated as

$$P(A) = \frac{\text{number of simple outcomes} \in A}{\text{total number of simple outcomes}}$$

Important Relationships Between Events

Relationship	Definition	Implies That...
disjoint or **mutually exclusive**	the events can never occur together	P(A and B) = 0, so **P(A or B) = P(A) + P(B)**

Knowing that events are disjoint can make things much easier, since otherwise P(A and B) can be difficult to find.

Relationship	Definition	Implies That...
complementary	the complement of event A (denoted A^C or $\overline{A}$) means "**not A**"; it consists of all simple outcomes in S that are not in A	$P(A) + P(A^C) = 1$ (*any event will either happen,* **or not**) thus, $P(A) = 1 - P(A^C)$; **$P(A^C) = 1 - P(A)$**

The law of complements is a useful tool, since it's **often easier to find the probability that an event does NOT occur.**

Relationship	Definition	Implies That...
independent	the occurrence of one event does not affect the probability of the other, and vice versa	P(A|B) = P(A), and P(B|A) = P(B), so **P(A and B) = P(A)P(B)**

Events are often assumed to be independent, particularly **repeated trials.**

Probability Rules

Rule	Formula
addition rule ("or")	P(A or B) = P(A) + P(B) - P(A and B) ***if A and B are disjoint,*** P(A or B) = P(A) + P(B)

Subtract P(A *and* B) so as not to count ***twice*** the elements of both A and B.

Rule	Formula
multiplication rule ("and")	P(A and B) = P(A)P(B|A) *equivalently,* P(A and B) = P(B)P(A|B) ***if A and B are independent,*** P(A and B) = P(A)P(B)

While it doesn't matter whether we "condition on A" (first) or "condition on B" (second), generally the information available will require one or the other.

Rule	Formula
conditional probability rule ("given that")	$P(A\|B) = \frac{P(A \text{ and } B)}{P(B)}$ $P(B\|A) = \frac{P(A \text{ and } B)}{P(A)}$

By multiplying both sides by P(B) or P(A), we see this is a rephrasing of the multiplication rule; conditional probabilities are often difficult to assess; an alternative way of thinking about "P(A|B)" is that it is *the proportion of elements in B that are* ***ALSO*** *in A.*

Rule	Formula
total probability rule	To find the probability of an event A, if the sample space is *partitioned* into several disjoint and exhaustive events $D_1, D_2, D_3, \ldots, D_k$, then, since A must occur along with one and only one of the D's: $P(A) = P(A \text{ and } D_1) + P(A \text{ and } D_2) + \ldots + P(A \text{ and } D_k)$ $\mathbf{= P(D_1)P(A\|D_1) + P(D_2)P(A\|D_2) + \ldots + P(D_k)P(A\|D_k)}$

The total probability rule may look complicated, but it isn't! (*see* sample problem 3a, next page).

Rule	Formula
Bayes' Theorem	With two events, A and B, using the total probability rule: $P(B\|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{P(A \text{ and } B)}{P(A \text{ and } B) P(A \text{ and } B^c)} = \frac{P(B)(A\|B)}{P(B)P(A\|B) + P(B^c)(A\|B^c)}$

Bayes' Theorem allows us to reverse the order of a conditional probability statement, and is *the only generally valid method!*

Probability Distributions

When some number is derived from a probability experiment, it is called a **random variable.**

Every random variable has a **probability distribution** that determines the probabilities of particular values.

For instance, when you roll a fair, six-sided die, the resulting number (X) is a random variable, with the following **discrete** **probability distribution:**

X	P(X)
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

In the table to the right, P(X) is called the **probability distribution function (pdf).**

Since each value of P(X) represents a probability, pdf's must follow the basic probability rules: P(X) must always be between 0 and 1, and all of the values P(X) sum to 1.

Other probability distributions are **continuous:** They do not assign specific probabilities to specific values, as above in the ***discrete case;*** instead, we can measure probabilities only over a **range** of values, using the area under the curve of a **probability density function.**

Much like data variables, we often measure the **mean ("expectation")** and **standard deviation** of random variables; if we can characterize a random variable as belonging to some major family (*see* table below), we can find the mean and standard deviation easily; in general, we have:

Type of Random Variable	General Formula for Mean	General Formula for Standard Deviation
discrete *(X takes some countable number of specific values)*	$\mu = E(X) = \sum X P(X)$	$\sigma = SD(X) = \sqrt{\sum X^2 P(X) - \mu^2}$
continuous *(X has uncountable possible values, and P(X) can be measured only over intervals)*	$\mu = E(X) = \int XP(X)dX$	$\sigma = SD(X) = \sqrt{\int X^2 P(X) dX - \mu^2}$

Fortunately, most useful continuous probability distributions do not require integration in practice; other formulas and tables are used.

Sample Problems & Solutions

1. Discrete random variable, *X*, follows the following probability distribution:

X	0	1	2	3	sums
P(X)	0.15	0.25	0.4	0.2	1 *(always)*
XP(X)	0	0.25	0.8	0.6	1.65=E(X)
X^2 P(X)	0	0.25	1.6	1.8	3.65

a. What is the expected value of X?

$$\mu = E + (X) = \sum XP(X) = \mathbf{1.65}$$

b. What is the standard deviation of X?

$$\sigma = SD(X) = \sqrt{\sum X^2 P(X) - \mu^2} =$$

$$\sqrt{3.65 - 1.65^2} = \sqrt{0.9275} = \mathbf{0.963}$$

Several Important Families of Discrete Probability Distributions

Name	Used When	Parameters	PDF	Mean	Standard Deviation
uniform	all outcomes are consecutive integers, and all are equally likely	*a* = minimum *b* = maximum	$P(X)=\frac{1}{b-a+1}$	$\frac{a+b}{2}$	$\sqrt{\frac{(b-a)^2}{12}}$
Not common in nature.					
binomial	some fixed number of independent trials with the same probability of a given event each time; X = total number of times the event occurs	*n* = fixed number of trials *p* = probability that the designated event occurs on a given trial	$P(X) = {}_nC_x p^x(1-p)^{n-x}$	*np*	$\sqrt{np(1-p)}$
Commonly used distribution; symmetric if *p* = 0.5; only valid values for X are $0 \le X \le n$.					
Poisson	events occur independently, at some average rate per interval of time/space; X = total number of times the event occurs	λ = mean number of events per interval	$P(X) = \frac{e^{-\lambda}\lambda^x}{x!}$	λ	λ
There is no upper limit on X for the Poisson distribution.					
geometric	a series of independent trials with the same probability of a given event; X = # of trials until the event occurs	*p* = probability that the event occurs on a given trial	$P(X) = (1-p)^{x-1}p$	$\frac{1}{p}$	$\sqrt{\frac{1-p}{p^2}}$
Since we only count trials until the event occurs the first time, there is no need to count the ${}_nC_x$ arrangements, as in the binomial.					
hyper-geometric	drawing samples from a finite population, with a categorical outcome X = # of elements in the sample that fall in the category of interest	*N* = population size *n* = sample size *K* = number in category in population	$P(X)=\frac{{}_KC_x \; {}_{N-K}C_{n-x}}{{}_NC_n}$	$n\left(\frac{K}{N}\right)$	$\sqrt{\frac{n(N-n)\left(\frac{K}{N}\right)\left(1-\frac{K}{N}\right)}{N-1}}$

Sample Problems & Solutions

1. A sock drawer contains nine black socks, six blue socks, and five white socks—none paired up; reach in and take two socks at random, *without replacement;* find the probability that...

There are 20 socks, total, in the drawer (9 + 6 + 5 = 20) before any are taken out; in situations like this, without any other information, we should assume that each sock is equally likely to be chosen.

a. ...both socks are black

P(both are black) = P(first is black AND second is black) = P(first is black)P(second is black | first is black)

$$=\frac{9}{20}\times\frac{8}{19}=\frac{9\times 8}{20\times 19}=\frac{72}{380}=\mathbf{0.189}$$

b. ...both socks are white

[Expect a smaller probability than in the preceding problem, as there are fewer white socks from which to choose!]

As above, we lose both one of the socks in the category, as well as one of the socks total, after selecting the first:

$$\frac{5}{20}\times\frac{4}{19}=\frac{5\times 4}{20\times 19}=\frac{20}{380}=\mathbf{0.053}$$

c. ...the two socks match *(i.e., that they are of the same color)*

There are only three colors of sock in the drawer:
P(match) = P(both black) + P(both blue) + P(both white)

$$=\frac{9}{20}\times\frac{8}{19}+\frac{6}{20}\times\frac{5}{19}+\frac{5}{20}\times\frac{4}{19}=\frac{122}{380}=\mathbf{0.321}$$

d. ...the socks **DO NOT** match

For the socks ***not*** to match, we could have the first black and the second blue, or the first blue and the second white...or a bunch of other possibilities, too; ***it is much safer, as well as easier, to use the rule for complements***—common sense dictates that the socks will either match or not match, so:
P(socks DO NOT match) = 1 – P(socks ***do*** match) – 1 – 0.321 = **0.690**

2. In a particular county, 88% of homes have air conditioning, 27% have a swimming pool, and 23% have both; what is the probability that one of these homes, chosen at random, has...

a. ...air conditioning **OR** a pool?

The given percentages can be taken as probabilities for these events, so we have: P(AC) = 0.88, P(pool) = 0.27 and P(AC and pool) = **0.23**

b. ...**NEITHER** air conditioning **NOR** a pool?

By the addition rule: P(AC or pool) = P(AC) + P(pool) – P(AC and pool) 0.88 + 0.27 – 0.23 = **0.92**

Upon examination ***of the event,*** this is the complement of the above event: P(neither AC nor pool) = P(no AC **AND** no pool) = 1 – P(AC ***or*** pool) = 1 – 0.92 = **0.08**

c. ...has a pool, ***given that*** it has air conditioning?

This is the same as asking, "What proportion of the homes with air conditioning also have pools?" Whenever we use the phrase ***"given that,"*** a **conditional probability** is indicated:

$$\text{P(pool | AC)} = \frac{P(pool\ and\ AC)}{P(AC)}=\frac{0.23}{0.88}=0.261$$

d. ...has air conditioning, ***given that*** it has a pool?

This probability is much greater, since more homes have air conditioning than pools.

[CAUTION! This is NOT the same as the preceding problem—now we're asked what proportion of homes that have pools ALSO have air conditioning.]

The event in the numerator is the same; what has changed is the **condition:**

$$\text{P(AC | pool)} = \frac{P(pool\ and\ AC)}{P(AC)}=\frac{0.23}{0.27}=\mathbf{0.852}$$

3. The TTC Corporation manufactures ceiling fans; each fan contains an electric motor, which TTC buys from one of three suppliers: 50% of their motors from supplier A, 40% from supplier B, and 10% from supplier C; of course, some of the motors they buy are defective—the defective rate is 6% for supplier A, 5% for supplier B, and 30% for supplier C; one of these motors is chosen at random; find the probability that...

We have here a bunch of statements of probability, and it's useful to list them explicitly; let events A, B, and C denote the supplier for a fan motor, and D denote that the motor is defective, then: P(A) = 0.5, P(B) = 0.4, and P(C) = 0.1

The information about defective rates provides conditional probabilities: P(D|A) = 0.06, P(D|B) = 0.05, and P(D|C) = 0.3

We can also note the complementary probabilities of a motor **not** being defective: $P(D^C|A) = 0.94$, $P(D^C|B) = 0.95$, and $P(D^C|C) = 0.7$

a. ...the motor is defective

To find the overall defective rate, we use the total probability rule, as a defective motor still had to come from supplier A, B, or C:
P(D) = P(A and D) + P(B and D) + P(C and D) = P(A)P(D|A) + P(B)P(D|B) + P(C)P(D|C) = (0.5)(0.06) + (0.4)(0.05) + (0.1)(0.3) = 0.03 + 0.02 + 0.03 = **0.08**

If 8% overall are defective, then 92% are not—that is, we can also conclude that $P(D^C) = 1 - P(D) = 1 - 0.08 = 0.92$

b. ...the motor came from supplier C, ***given that*** it is defective

This is like asking, "What proportion of the defectives come from supplier C?" Denote this probability as P(C|D); we began with P(D|C) (among other probabilities)—we are effectively using **Bayes' Theorem** to reverse the order; however, we already have P(D), so:

$$\text{P(C|D)} = \frac{P(C\ and\ D)}{P(D)}=\frac{0.03}{0.08}=\mathbf{0.375}$$

PROBABILITY *(continued)*

Continuous Probability Distribution

Computer software or printed tables are usually used to compute probabilities for continuous random variables, but some important families include:

Name	Denoted	Parameters	Properties
normal (Gaussian)	X (or some other letter)	μ = mean σ = standard deviation	symmetric, unbounded, bell-shaped; arises commonly in nature and in statistics, as a result of the **central limit theorem**
Many other distributions approach the normal as n (or some other parameter, such as λ or df) increases.			
standard normal	Z	μ = mean = **0** σ = standard deviation = **1**	a special variant of normal, with **μ = 0** and **σ = 1;** *represented in "Z tables"*
Used for inference about proportions; the **cumulative probability** is provided in Z tables: For a particular value z, the **cumulative probability** is $\Phi(z) = P(Z < z)$; i.e., *the area under the density curve to the left of z.*			
student's t	t	*df = degrees of freedom*	similar in shape to normal **μ = 0 (always!)**
Used for inference about means.			
chi-square	χ^2	*df = degrees of freedom*	not symmetric (skewed right)
Used for inferences about categorical distributions.			

Sample Problems & Solutions

1. For a standard normal random variable Z, find $P(Z < 1.5)$.

Since, by definition, the values from the standard normal table are $\Phi(z) – P(Z < z)$ ... $P(Z < 1.5) = \Phi(1.5) =$ **0.9332**

2. For a t distribution with $df = 20$, which critical value of t has an area of 0.05 in the right tail?

A t table generally provides the tail area, rather than the cumulative probability, as given in standard normal tables; with the **row = df = 20,** and the **column = tail area = 0.05,** a t table produces the value of **1.725**

3. The heights of military recruits follow a normal distribution with a mean of 70 inches and a standard deviation of 4 inches; find the probability that a randomly chosen recruit is...

a. shorter than 60 inches

First, we must transform values of the variable (height) to the standard normal distribution, by taking z scores; here:

$$z = \frac{x-\mu}{\sigma} = \frac{60-70}{4} = \frac{-10}{4} = \mathbf{-2.5}$$

⚠ Since we want the **"less than"** probability, the solution comes directly from the standard normal z table:

$P(X < 60) = P(Z < -2.5) = \Phi(-2.5) =$ **0.0062**

b. taller than 72 inches

First, the z score: $z = \frac{x-\mu}{\sigma} = \frac{72-70}{4} = \frac{2}{4} = \mathbf{0.5}$

⚠ Since this is a **"greater than"** probability, subtract the cumulative probability from 1:

$P(X > 72) = P(Z > 0.5) = 1 - \Phi(0.5) = 1 - 0.6915 =$ **0.3085**

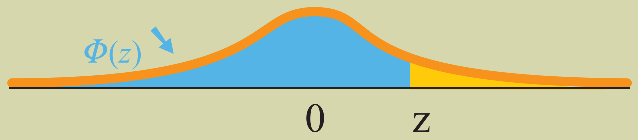
c. between 64 and 76 inches tall

In this case, there are two boundaries: The only way to find the area under the curve between them is to find the cumulative probabilities for each, and then to subtract; this entails finding z scores for both X = 64 <u>and</u> X = 76:

$$z = \frac{x-\mu}{\sigma} = \frac{64-70}{4} = \frac{-6}{4} = \mathbf{-1.5} \text{ and } z = \frac{x-\mu}{\sigma} = \frac{76-70}{4} = \frac{6}{4} = \mathbf{1.5}$$

Now:
$P(64 < x < 76) = P(-1.5 < Z < 1.5) = \Phi(1.5) - \Phi(-1.5) = 0.9332 - 0.0668 =$ **0.8664**

$$\Phi(z) = P(Z<z)$$



SAMPLING DISTRIBUTIONS

statistic	expected value	standard error
sample mean	μ	$\frac{\sigma}{\sqrt{n}}$
...if $n \geq 30$, <u>**or**</u> if the population distribution is normal		
sample proportion	p	$\sqrt{\frac{p(1-p)}{n}}$
...if $np \geq 15$ <u>**and**</u> $n(1 - p) \geq 15$		

Because sample statistics are derived from random samples, they are random. The probability distribution of a statistic is called its sampling distribution.

Due to the central limit theorem, some important statistics have sampling distributions that approach a normal distribution as the sample size increases (these are listed in the table at right).

Knowing the expected value and standard error allows us to find probabilities; then, in turn, we can use the properties of these sampling distributions to make inferences about the parameter values when we *do not know them,* as in real-world applications.

Sample Problems & Solutions

1. 60% of the registered voters in a large district plan to vote in favor of a referendum; a random sample of 340 of these voters is selected.

a. What is the expected value of the sample proportion?

$$E(p) = p = \mathbf{0.6}$$

b. What is the standard error of the sample proportion?

$$SE(p) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.6(1-0.6)}{340}} = \mathbf{0.0266}$$

c. What is the probability that the sample proportion is between 55% and 65%?

First, find the z scores for those proportions:

$$z = \frac{p(p)}{SE(p)} = \frac{0.55-0.6}{0.0266} = \frac{-0.05}{0.0266} = \mathbf{-1.88} \text{ \underline{and}}$$

$$z = \frac{p(p)}{SE(p)} = \frac{0.65-0.6}{0.0266} = \frac{0.05}{0.0266} = \mathbf{1.88}$$

Now,
$P(0.55)\,\hat{p}\,(0.65) = P - (1.88)\,Z\,(1.88)$
$= \Phi(1.88) - \Phi(-1.88) = 0.9699 - 0.0301 =$ **0.9398**

2. The standard deviation of the weight of cattle in a certain herd is 160 pounds, but the mean is unknown; a random sample of size 100 is chosen.

a. Compute the standard error of the sample mean:

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}} = \frac{\mathbf{160}}{\sqrt{\mathbf{100}}} = \mathbf{16}\ \textbf{\textit{lbs.}}$$

b. For an **individual animal** in this herd, what is the probability of a weight within 40 lbs. of the population mean?

Since this problem refers to a **single observation,** *not the sample mean*, we <u>**use the standard deviation**</u>, not the standard error.

⚠ Not knowing the value of μ, we can only express the boundaries for "within 40 lbs. of the mean" as $X = \mu + 40$ and $X = \mu - 40$

We can still compute z scores:

$$z = \frac{x-\mu}{\sigma} = \frac{\mu+40-\mu}{160} = \frac{40}{160} = \mathbf{0.25} \text{ \underline{and}}$$

$$z = \frac{x-\mu}{\sigma} = \frac{\mu-40-\mu}{160} = \frac{-40}{160} = \mathbf{-0.25}$$

That is, "within 40 lbs. of the mean" is the same as **within 0.25 standard deviation.**

We find the probability:
$P(-0.25 < Z < 0.25) = \Phi(0.25) - \Phi(-0.25) = 0.5987 - 0.4013 =$ **0.1974**

c. What is the probability that the sample mean falls within 40 lbs. of the population mean?

Even though we don't know the population mean, the z score formula will allow us to find this probability.

⚠ Since this is the sample **mean,** we must use the standard error of **16 lbs.,** rather than the standard deviation, in computing the z scores:

$$z = \frac{x-\mu}{SE(\bar{X})} = \frac{\mu+40-\mu}{16} = \frac{40}{16} = \mathbf{2.5} \text{ \underline{and}}$$

$$z = \frac{x-\mu}{SE(\bar{X})} = \frac{\mu-40-\mu}{16} = \frac{-40}{16} = \mathbf{-2.5}$$

Now:
$P(-2.5 < Z < 2.5) = \Phi(2.5) - \Phi(-2.5) = 0.9938 - 0.0062 =$ **0.9876**

⚠ This probability is dramatically higher than the probability for an individual head of cattle!

STATISTICAL INFERENCE

Null and alternative hypotheses have the following very important properties:

When we want to draw conclusions about a population using data from a sample, we use some method of **statistical inference.**

A **hypothesis test** is a procedure by which claims about populations *(hypotheses)* are evaluated on the basis of sample statistics.

The procedure begins with a **null hypothesis (H_0)** and an *alternative* (or "research") *hypothesis (H_1)*; if the sample data are too unusual, assuming H_0 to be true, then H_0 is rejected in favor of H_1; otherwise, we fail to reject the null hypothesis, and thereby fail to support the alternatives.

the null hypothesis (H_0)	the alternative hypothesis (H_1 or H_a)
is assumed true for the purpose of carrying out the hypothesis test	is supported only by carrying out the test, if the null hypothesis can be rejected
ALWAYS provides a specific value for the parameter, its ***"null value"***; always contains "="	**NEVER** provides a specific value for the parameter; instead, contains ">" (**right-tailed**), "<" (**left-tailed**), or "≠" (**two-tailed**)
the **null value** implies a *specific sampling* distribution for the **test statistic**	without any specific value for the parameter of interest, the sampling distribution is unknown
can be rejected—or not rejected—but NEVER ***supported***	can be supported (by rejecting the null)—or not supported (by failing to reject the null)—**but NEVER *rejected***

The tail(s) of the hypothesis test are determined by the alternative hypothesis (H_1)—**this is one of the <u>most important attributes of the test</u>, regardless of which method is used.**

Steps for Carrying Out a Hypothesis Test

There are two major methods for carrying out a hypothesis test: the **traditional approach** (or *fixed significance*) and the ***p*-value approach** (*observed significance*); the following table lists the steps for each approach:

p-value approach	traditional approach
formulate **null** and **alternative hypotheses**	formulate a null and an alternative hypothesis
observe sample data	determine rejection region(s) based on the level of significance and the tail(s) of the test
compute a **test statistic** from sample data	observe sample data
compute the ***p*-value** from the test statistic	compute the **test statistic** from sample data
reject the null hypothesis (supporting the alternative) at a significance level α, if the ***p*-value** $\leq \alpha$; otherwise, fail to reject the null hypothesis	reject the null hypothesis (supporting the alternative) at the significance level, **if the test statistic falls in the rejection region;** otherwise, fail to reject the null hypothesis

With the *p*-value approach, the final decision is made by comparing probabilities, whereas with the traditional approach, the decision is made by comparing values of random variables; because there is a one-to-one correspondence between the values of the random variables and the probabilities, **the two methods will always yield consistent results;** we can convert between the two using the following simple (but important!) rule:

reject the null hypothesis (H_0) at significance level α $\rightleftarrows$ ***p*-value** $\leq \alpha$

Test Statistics

Parameter	Test Statistic	Distribution Under H_0	Assumptions
population proportion	$Z = \dfrac{\hat{p} - p_0}{SE(\hat{p})}$	standard normal Z	$np \geq 15$ <u>**and**</u> $n(1-p) \geq 15$
population mean	$t = \dfrac{\bar{x} - \mu_0}{SE(\bar{x})}$	t distribution with $df = n - 1$	$n \geq 30$, **or** the population distribution is normal

Since the t distribution approaches the standard normal Z, many teachers and texts advise that it's OK to use Z if n is sufficiently large.

Parameter	Test Statistic	Distribution Under H_0	Assumptions
difference of proportions (independent samples)			$np \geq 15$ <u>**and**</u> $n(1-p) \geq 15$
test for independence (categorical data)	$\chi^2 = \sum \dfrac{(O-E)^2}{E}$	χ^2 distribution with $df = (r-1)(c-1)$; r = # of rows; c = # of columns	χ^2 tests for categorical data assume that the expected counts (E) in each cell are **at least 5** under the null hypothesis
multinomial goodness-of-fit (categorical data)		χ^2 distribution with $df = k - 1$ <u>**and**</u> k = # of categories	

χ^2 tests for categorical data **do not have directional alternative hypotheses;** rejection regions are always in the right tail.

Sample Problems & Solutions

In each of the following cases, formulate hypotheses to test the claim; indicate which hypothesis represents the claim.

1. The manager of a bank claims that the average waiting time for customers is less than two minutes.

Since the claim refers to the average, this is a test for μ.

As a "less than" claim, it is represented by H_0, and the hypothesis test is: **H_0: $\mu = 2$, vs. H_1: $\mu < 2$**

(left-tailed)

2. Your friend says that a coin you are tossing is not fair.

A fair coin is one that shows heads 50% of the time; the friend states that the coin is NOT fair.

This is an **H_1** claim: **H_0: $p = 0.5$, vs. H_1: $p \neq 0.5$**

(two-tailed)

3. A highway patrolman claims that the average speed of cars on a highway is at most 70 mph.

The claim directly refers to the average; since this is an "at most" claim, it is represented by **H_0**.

The hypothesis test is: **H_0: $\mu = 70$, vs. H_1: $\mu > 70$**

(right-tailed)

4. A motorist claims that more than 80% of the cars on a highway travel at a speed exceeding 70 mph.

Since the claim is really about a proportion–don't be fooled by the "70 mph!"—the hypotheses refer to *p*.

As the motorist makes a "more than" claim, it is the null hypothesis, **H_0.**

H_0: $p = 0.8$, vs. H_1: $p > 0.8$

(right-tailed)

5. The manager of a snack-food factory states that the average weight of a bag of their potato chips is exactly 5 oz. (no more, no less).

This is an "is exactly" claim that refers to the average; thus, the claim is **H_0.**

The test is: **H_0: $\mu = 5$, vs. H_1: $\mu \neq 5$**

(two-tailed)

Formulating Hypotheses

if claim consists of...	it is represented by...
"...is not equal to..."	alternative hypothesis (H_1)
and the hypothesis test is two-tailed ≠	
"...is less than..."	alternative hypothesis (H_1)
and the hypothesis test is left-tailed <	
"...is greater than..."	alternative hypothesis (H_1)
and the hypothesis test is right-tailed >	
"...is equal to..."/"...is exactly..."	null hypothesis (H_0)
and the hypothesis test is two-tailed ≠	
"...is at least..."	null hypothesis (H_0)
and the hypothesis test is left-tailed <	
"...is at most..."	null hypothesis (H_0)
and the hypothesis test is right-tailed >	

Errors in Inference

Decision	Reality	
	H_0 true	H_0 false
reject H_0 (supporting H_1)	**type I error** $P(\text{reject } H_0 \mid H_0 \text{ true}) = \alpha$ = level of **significance**	**correct inference** $P(\text{reject } H_0 \mid H_0 \text{ false}) = \mathbf{1-\beta}$ = **power**
⚠ When the null hypothesis (H_0) is rejected, we can support the alternative hypothesis (H_1). This is a substantive finding: We have sufficient evidence that H_0 is not correct.		
fail to reject H_0 (failing to support H_1)	**correct inference** $P(\text{fail to reject } H_0 \mid H_0 \text{ false}) =$ **$1-\alpha$ = level of confidence**	**type II error** $P(\text{fail to reject } H_0 \mid H_0 \text{ true}) = \beta$
⚠ If H_0 is not rejected, then we cannot support H_1 either; this is **NOT** a substantive finding: We have failed to find evidence against H_0, **but have not "confirmed" or "proved" it to be true!**		
notes	Under the null hypothesis, *we have a specific value for the parameter* This determines a specific sampling distribution, so that α and $1-\alpha$ can be precisely determined.	If the null hypothesis is false, *there is no specific value for the parameter* Thus, we can only estimate β and $1-\beta$ by making some alternative assumption about the parameter.
⚠ It is important to note that these probabilities are conditioned on **reality,** rather than the *decision*. That is, given that H_0 is true, α is the probability of rejecting **H_0; it is NOT the probability that H_0 is true, given that it has been rejected!**		

Sample Problems & Solutions

1. In some hypothesis tests, the null hypothesis is rejected; if an error has been made, which kind of error is it?

⚠ The only error of inference in which the null hypothesis is rejected is a **type I error.**

2. A researcher conducts a hypothesis test at a significance level of 0.05, and computer software produces a *p*-value of 0.0912; unknown to the researcher, the null hypothesis is really false—what is her decision…Is it some type of error?

First, consider her decision: She will reject or fail to reject the null hypothesis; we have no test statistic, only a *p*-value.

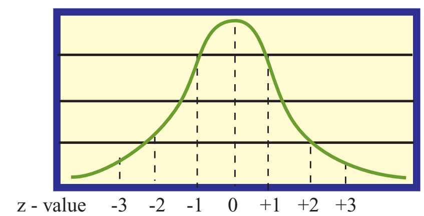
⚠ But, since the *p*-value is less than the significance level, α, **H_0 is rejected;** but also, since H_0 is *false*, this is a **type II error.**

Finding Rejection Regions & P-Values

Tail(s) of Hypothesis Test	Rejection Region	*P*-Value
< left-tailed	values of the test statistic **less than** some critical value with area α in the **left** tail	area under the density curve to the left of the test statistic
> right-tailed	values of the test statistic **greater than** some critical value with area α in the **right** tail	area under the density curve to the right of the test statistic
≠ two-tailed	values of the test statistic **less than** some critical value with area α in the **left** tail, **or** *greater than* some critical value with area α in the *right* tail	**double** the tail area under the curve *away from the test statistic*

Percentage Cumulative Distribution

for selected z values under a normal curve



Sample Problems & Solutions

1. At an aquaculture facility, a large number of eels are kept in a tank; they die independently of each other at an average rate of 2.5 eels per day.

a. Which distribution is appropriate?

Since the events are independent, and we're given an **average rate per fixed interval,** a Poisson distribution can be used, with parameter: $\lambda = 2.5$

b. Find the probability that exactly two eels die in a given day:

Find P(X) for X = 2

$$P(2) = \frac{e^{-2.5}2.5^2}{2!} = \mathbf{0.1283}$$

c. What is the probability that at least one eel dies in the span of one day?

Since the Poisson distribution has **no maximum**, there is no alternative but to use the law of complements: P(at least one dies)= 1– P(none at all die) =

$$1 - P(0) = 1 - \frac{e^{-2.5}2.5^0}{0!} = 1 - e^{-2.5} = 1 - 0.0821 = \mathbf{0.9179}$$

HARD d. Compute the probability that at least one eel dies in the span of 12 hours:

⚠ This is harder, since the duration of the interval has changed; but, we can scale the Poisson parameter λ proportionally: If the average rate is 2.5 eels per day, then the rate is 1.25 (half as many) per half-day; thus:

$$1 - P(0) = 1 - \frac{e^{-1.25}1.25^0}{0!} = 1 - e^{-1.25} = 1 - 0.2865 = \mathbf{0.7135}$$

2. A cat is hunting some mice; every time she pounces at a mouse, she has a 20% chance of catching the mouse, but will stop hunting as soon as she catches one.

a. Which distribution is appropriate?

As there is a fixed probability of the event, but the experiment will be repeated until the event occurs, a geometric distribution can be used, with parameter p = **0.2**

EASY **b.** What is the probability that she'll catch a mouse on her first attempt? With a 20% chance of success each time, the probability of succeeding the first time is simply 0.2

We can also use the geometric pdf, with x=1: $P(1) = (1-0.2)^{1-1}(0.2) = \mathbf{0.2}$

c. What is the probability that she'll catch a mouse on her third attempt?

The first success occurring on the third trial means **x = 3:** $P(3) = (1-0.2)^{3-1}(0.2) = (0.8)^2(0.2) = \mathbf{0.128}$

d. How many times is she expected to pounce until she succeeds?

$$E(X) = \frac{1}{p} = \frac{1}{0.2} = \mathbf{5}$$

3. John is playing darts; each time he throws a dart, he has an 8% chance of hitting a bull's-eye, independently of the result for any other dart thrown; he throws a total of five darts.

a. Which distribution is appropriate?

With a constant probability of success, and a fixed number of independent events, the total number of successes follows a **binomial distribution,** with parameters: **n = 5, p = 0.08**

b. How many bull's-eyes is John expected to hit? $E(X) = np = 5(0.08) = \mathbf{0.4}$

c. What is the probability that he hits exactly two bull's-eyes?

$x = 2$: $P(X) = {}_5C_2\, 0.08^2 (1-0.08)^{5-2} = (10)(0.0064)(0.92)^3 = \mathbf{0.0498}$

d. What is the probability that he hits at least one bull's-eye?

As always, P(at least one) = 1 – P(none at all) $= 1 - P(0) = 1 - {}_5C_0\, 0.08^0(1-0.08)^{5-0} = 1 - 0.92^5 = 1 - 0.6591 = \mathbf{0.3409}$

U.S. $5.95

Customer Hotline # 1.800.230.9522

AUTHOR: Stephen V. Kizlik, Ph.D.

ISBN-13: 978-142320969-0
ISBN-10: 142320969-9

50595

9 781423 209690

6 54614 00969 2