SOLUTIONS

MAT 167: Statistics

Final Exam

Instructor: Anthony Tanbakuchi

Fall 2007

Name: ______________________________

Computer / Seat Number: ____________

Multiple choice part: fill in answer on the scan form. Do not attach work for this section (no partial credit is awarded).
Written part: Write all final answers on the exam. Actual work should be attached so partial credit can be given.

No books, notes, or friends. You may use the attached equation sheet, R, and a calculator. No other materials. If you choose to use R, copy and paste your work into a word document labeling the question number it corresponds to. When you are done with the test print out the document. Be sure to save often on a memory stick just in case. Using any other program or having any other documents open on the computer will constitute cheating.

You have until the end of class to finish the exam, manage your time wisely.
If something is unclear quietly come up and ask me.
If the question is legitimate I will inform the whole class.

Express all final answers to 3 significant digits. Probabilities should be given as a decimal number unless a percent is requested. Circle final answers, ambiguous or multiple answers will not be accepted. Show steps where appropriate.

The exam consists of 9 questions for a total of 36 points on 15 pages.

This Exam is being given under the guidelines of our institution's **Code of Academic Ethics**. You are expected to respect those guidelines.

Points Earned: ____________ **out of 36 total points**

Exam Score: ____________

1. Given the following data from a random sample:

   21 11 21 14 20 26 17 220

   (a) (1 point) Find the mean of the data.

   Solution:

   ```
   > mean(x)
   [1] 43.75
   ```

   (b) (1 point) Find the median of the data.

   Solution:

   ```
   > median(x)
   [1] 20.5
   ```

   (c) (1 point) Since the mean and the median are not equal, what does this indicate about the data?

   Solution: The data is skewed.

   (d) (1 point) Which measure of center (mean or median) do you think is better for describing this data and **why**?

   Solution: (Answers may vary somewhat) This data set contains one outlier so the **median** would be a better measure of the center since it is less susceptible to outliers.

   (e) (1 point) Find the standard deviation of the data.

   Solution:

   ```
   > sd(x)
   [1] 71.36576
   ```

2. In regards to $\bar{x}$ and the Central Limit Theorem:

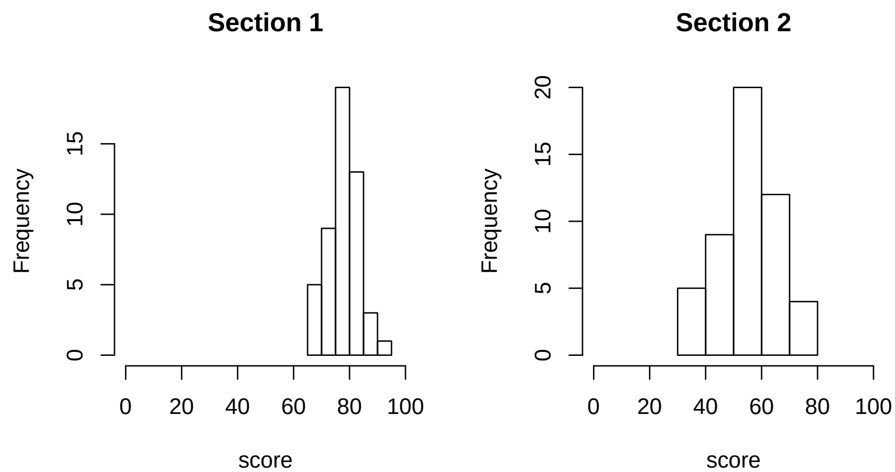
   (a) (1 point) What are the two conditions under which the CLT applies?

   Solution: Either (1) x has a normal distribution or (2) $n > 30$.

   (b) (1 point) If the conditions are met, what does the CLT state about $\bar{x}$?

   Solution: The sampling distribution of $\bar{x}$ will have a normal distribution.

3. The following are histograms of student scores on the same exam for two different sections of a statistics class.



(a) (1 point) Which section had a higher mean score?

Solution: Section 1

(b) (1 point) Which section had a larger standard deviation?

Solution: Section 2

4. The following questions regard hypothesis testing in general.

(a) (1 point) When we conduct a hypothesis test, we assume something is true and calculate the probability of observing the sample data under this assumption. What do we assume is true?

Solution: We assume the null hypothesis H_0 is true.

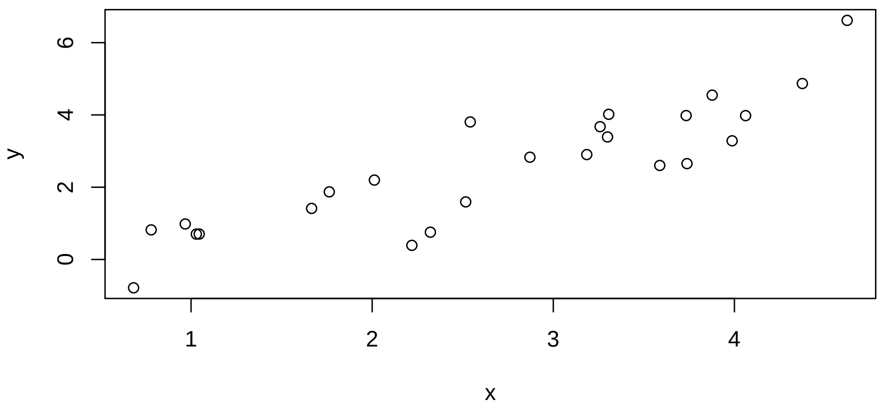
(b) (1 point) If you are using a hypothesis test to make a decision where the effect of a Type I error may negatively effect human lives, should you increase or decrease the significance level α used in making the decision?

Solution: You should **decrease** the significance level α to reduce the probability of making a Type I error.

(c) (1 point) The 1-Way ANOVA is a many sample generalization of what two sample test?

Solution: The two sample t-test.

5. Use the following plot of paired x-y data and the computer analysis output for the following questions, assume that the linear correlation coefficient is statistically significant.




```
> mean(x)
[1] 2.697902
> mean(y)
[1] 2.552376
> results = lm(y ~ x)
> results
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)          x
      -0.82       1.25
```

(a) (1 point) Which linear correlation coefficient below best matches the plot?

10, 1, 0.9, 0.09, 0, -0.09, -0.9, -1, -10

> **Solution:**
> $r = 0.9$.

(b) (1 point) Based on your above answer, what percent of variation in y is explained by x?

> **Solution:** r^2 describes the percent of variation in y explained by x. Thus, $r^2 = 81\%$.

(c) (1 point) What is the linear equation that models the data?

> **Solution:**
> $$\hat{y} = -0.82 + 1.25x$$

(d) (1 point) Using your previous answers, what is your best point estimate for y if $x = 4$?

Solution: Plug in the value of x into your equation above:

```
> out = predict(results, newdata = data.frame(x = 4), int = "pred")
> signif(out[1], 3)
[1] 4.18
```

6. You are a crime scene investigator trying to match the lead content of bullet fragments found at a crime scene to the lead content of a box of bullets found with a suspect. To simplify this question, assume that the instrument you use gives you one measurement per fragment in grams/cm^3. Assume that you have 5 measurements from fragments found at the crime scene and 7 measurements from bullets found with the suspect.
   (a) (1 point) What type of hypothesis test will you use?

   Solution: Two sample hypothesis test for equality of means. (Two sample t-test.)

   (b) (1 point) What are H_0 and H_a? Write them both mathematically and in words.

   Solution: $H_0 : \mu_1 = \mu_2$, the mean lead content in the bullets used at the crime scene is the same as the mean lead content in the bullets found in the suspects car.
   $H_a : \mu_1 \neq \mu_2$, the mean lead content does not match between the crime scene's bullets and the suspect's bullets.

   (c) (1 point) You run the analysis and the p-value is 0.0001 and $\alpha = 0.001$, and $\beta = 0.9$. If you **reject** H_0, what is the probability that you made the wrong decision in this case?

   Solution: The probability of a Type I error is the p-value$= 0.0001$.

   (d) (1 point) You run the analysis and the p-value is 0.9, $\alpha = 0.001$ and $\beta = 0.01$. If you **fail to reject** H_0, What is the probability that you made the wrong decision in this case?

   Solution: The probability of a Type II error is the $\beta = 0.01$.

   (e) (1 point) Under what conditions could an expert witness give the following statement based solely on lead bullet statistical evidence:
   "The bullet fragments must have come from the same box or from another box that would have been made by the same company on the same day."

   Solution: The statement is NOT a valid conclusion that a statistical test would support. It would never be possible to say that fragments came from a specific box or day, the variability in too great to make such a statement. Moreover, "must" indicates no possibility of error, and no statistical test can ever prove a hypothesis. It can only show that it's likelihood is very small. It cannot say the likelihood is exactly 0.

7. The clothing manufacturer's association (CMA) publishes data that manufacture's use to determine what sizes of clothing they should make. As mentioned before, the CMA states that

men's waists are normally distributed with $\mu = 35$ in and $\sigma = 2.3$ in. You believe that the mean waist size of men is actually larger than 35.

(a) (1 point) You would like to conduct a study to estimate (at the 90% confidence level) the mean waist size of men with a margin of error of 1 in. Assuming that the standard deviation of waist sizes is $\sigma = 2.3$ in, what sample size should you use for this study?

Solution:
Using the equation

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2 \tag{1}$$

```
> E = 1
> sigma = 2.3
> alpha = 0.1
> z.critical = qnorm(1 - alpha/2)
> z.critical
[1] 1.644854
> n = (z.critical * sigma/E)^2
> n
[1] 14.31232
> ceiling(n)
[1] 15
```

(b) (1 point) A study was conducted (and they ignored your recommendation of sample size!) of 5 randomly selected men and the following waist sizes were measured:

37.3 43.5 36.3 34 35.4

Construct a 90% confidence interval for the true population mean waist size using the above data. (**Assume σ is unknown.**)

Solution:
Need to find E in

$$CI = \bar{x} \pm E \tag{2}$$
$$= \bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \tag{3}$$

```
> x
[1] 37.3 43.5 36.3 34.0 35.4
> alpha = 0.1
> n = length(x)
> x.bar = mean(x)
> x.bar
[1] 37.3
> s = sd(x)
> s
```

```
[1] 3.672193
> std.err = s/sqrt(n)
> std.err
[1] 1.642255
> t.crit = qt(1 - alpha/2, df = n - 1)
> t.crit
[1] 2.131847
> E = t.crit * std.err
> E
[1] 3.501035
```

The confidence interval is: 37.3 ± 3.5 or $(33.8, 40.8)$

(c) (1 point) Can you reject the claim that the mean waist size of men is 35 in based on the confidence interval that you constructed above?

Solution: No, we cannot reject the claim that the mean waist size is 35 in because the confidence interval contains 35.

8. A group supporting Hillary Clinton who sees Mitt Romney as her strongest republican competitor makes the statement that "Hillary has more support in the democratic party than Mitt Romney has in the republican party". After a little investigating you find out the political group did a random survey of 500 democratic voters and found 170 supported Hillary. They conducted a second survey of of 420 republican voters and found 135 supported Romney. Does this data support the claim that the proportion of democrats who support Hillary is greater than the proportion of republicans who support Romney?[1]

(a) (1 point) What type of hypothesis test will you use?

Solution: Use a two sample proportion test.

(b) (1 point) What are the test's requirements?

Solution: (1) Simple random samples, (2) normal approx. to binomial applies to both samples (np and $nq \geq 5$ for both groups).

(c) (1 point) Are the requirements satisfied? **State how they are satisfied.**

Solution: Yes. Simple random samples used, and normal approximation to binomial is satisfied.

(d) (1 point) What are the hypothesis H_0 and H_a?

Solution: Let Hillary be proportion 1.
$H_0 : p_1 = p_2$, $H_a : p_1 > p_2$

[1] The sample data in this problem is fictitious and is not an endearment of any candidate.

(e) (1 point) What α will you use?

Solution: $\alpha = 0.05$

(f) (1 point) Conduct the hypothesis test. What is the p-value?

Solution:

```
> x = c(170, 135)
> n = c(500, 420)
> res = prop.test(x, n, alternative = "greater")
> res
        2-sample test for equality of proportions with continuity correctio

data:  x out of n
X-squared = 0.2764, df = 1, p-value = 0.2995
alternative hypothesis: greater
95 percent confidence interval:
 -0.03479792  1.00000000
sample estimates:
   prop 1    prop 2
0.3400000 0.3214286
```

The p-value is 0.3.

(g) (1 point) What is your formal decision?

Solution: Since p-val $\nleq \alpha$, fail to reject H_0.

(h) (1 point) State your final conclusion in words.

Solution: The sample data does not support the claim that proportion of democrats who support Hillary is greater than the proportion of republicans who support Romney.

(i) (1 point) Assume that you failed to reject H_0. Hillary's political party truly believes that they do have more support. If they were to re-run the study, what should they change to increase their chances of being able to statistically support their claim?

Solution: Increase the sample size.

9. Chest deceleration data are given below. A researcher wants to test the claim that vehicle size has an effect on the mean chest deceleration at the 0.05 significance level.

vehicle size	chest deceleration (g)
Subcompact:	55, 47, 59, 49, 42
Compact:	57, 57, 46, 54, 51
Midsize:	45, 53, 49, 51, 46
Full-size:	44, 45, 39, 58, 44

(a) (1 point) What type of hypothesis test (of those discussed in class) should you use?

> **Solution:** 1-Way ANOVA

(b) (1 point) What is the null hypothesis for this test?

> **Solution:** H_0 : mean chest deceleration is the same for all weight classes.

(c) (1 point) If you analyze the data and your p-value is 0.2, what would your conclusion be?

> **Solution:** The sample data does not support the claim that vehicle size has an effect on the mean chest deceleration.

Fill in answers on the scan form for this part of the test.

MULTIPLE CHOICE. Choose the one alternative that best completes the statement or answers the question.

Determine whether the given value is a statistic or a parameter.

1) A health and fitness club surveys 40 randomly selected members and found that the average weight of those questioned is 157 lb. 1) ________

A) Parameter B) Statistic

Identify which of these types of sampling is used: random, stratified, systematic, cluster, convenience.

2) The name of each contestant is written on a separate card, the cards are placed in a bag, and three names are picked from the bag. 2) ________

A) Systematic
B) Convenience
C) Random
D) Cluster
E) Stratified

Construct the cumulative frequency distribution that corresponds to the given frequency distribution.

3) 3) ________

Height (inches)	Frequency
69.0 – 71.9	17
72.0 – 74.9	21
75.0 – 77.9	21
78.0 – 80.9	18
81.0 – 83.9	3

A)

Height (inches)	Cumulative Frequency
69.0 – 71.9	38
72.0 – 74.9	59
75.0 – 77.9	77
78.0 – 80.9	80
81.0 – 83.9	83

B)

Height (inches)	Cumulative Frequency
69.0 – 71.9	0.212
72.0 – 74.9	0.263
75.0 – 77.9	0.263
78.0 – 80.9	0.225
81.0 – 83.9	0.037

C)

Height (inches)	Cumulative Frequency
69.0 – 71.9	17
72.0 – 74.9	38
75.0 – 77.9	59
78.0 – 80.9	75
81.0 – 83.9	80

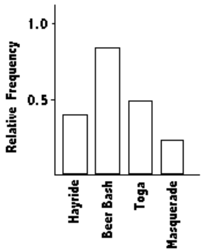
D)

Height (inches)	Cumulative Frequency
69.0 – 71.9	17
72.0 – 74.9	38
75.0 – 77.9	59
78.0 – 80.9	77
81.0 – 83.9	80

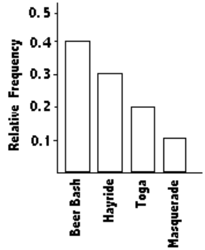
Solve the problem.

4) The Kappa Iota Sigma Fraternity polled its members on the weekend party theme. The vote was as follows: six for toga, four for hayride, eight for beer bash, and two for masquerade. Display the vote count in a Pareto chart. 4) ________

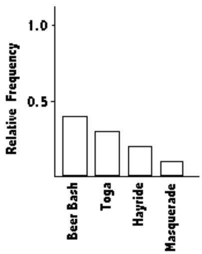
A)



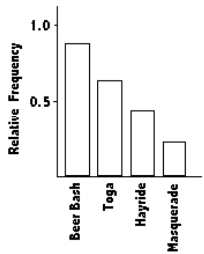
B)



C)



D)



Find the z–score corresponding to the given value and use the z–score to determine whether the value is unusual. Consider a score to be unusual if its z–score is less than –2.00 or greater than 2.00. Round the z–score to the nearest tenth if necessary.

5) A time for the 100 meter sprint of 13.7 seconds at a school where the mean time for the 100 meter sprint is 17.5 seconds and the standard deviation is 2.1 seconds. 5) ________

A) –3.8; unusual
B) –1.8; not unusual
C) –1.8; unusual
D) 1.8; not unusual

Find the mode(s) for the given sample data.

6) 20, 27, 46, 27, 49, 27, 49 6) ________

A) 46
B) 27
C) 35
D) 49

Provide a written description of the complement of the given event.

7) When 100 engines are shipped, none of them are defective. 7) ________

A) All of the engines are defective.
B) At least one of the engines is defective.
C) None of the engines are defective.

Find the indicated probability.

8) A study conducted at a certain college shows that 65% of the school's graduates find a job in their chosen field within a year after graduation. Find the probability that among 5 randomly selected graduates, at least one finds a job in his or her chosen field within a year of graduating. 8) ________

A) 0.650 B) 0.995 C) 0.200 D) 0.884

9) The table below shows the soft drinks preferences of people in three age groups. 9) ________

	cola	root beer	lemon–lime
under 21 years of age	40	25	20
between 21 and 40	35	20	30
over 40 years of age	20	30	35

If one of the 255 subjects is randomly selected, find the probability that the person is over 40 years of age given that they drink root beer.

A) $\frac{6}{17}$ B) $\frac{2}{5}$

C) $\frac{5}{17}$ D) None of the above is correct.

10) A sample of 4 different calculators is randomly selected from a group containing 48 that are defective and 23 that have no defects. What is the probability that all four of the calculators selected are defective? 10) ________

A) 0.2089 B) 21.9740 C) 0.0527 D) 0.2003

Find the standard deviation, σ, for the binomial distribution which has the stated values of n and p. Round your answer to the nearest hundredth.

11) $n = 2320$; $p = .63$ 11) ________

A) $\sigma = 26.52$ B) $\sigma = 20.84$ C) $\sigma = 23.25$ D) $\sigma = 27.37$

Find the indicated probability.

12) The incomes of trainees at a local mill are normally distributed with a mean of $1100 and a standard deviation $150. What percentage of trainees earn less than $900 a month? 12) ________

A) 9.18% B) 35.31% C) 90.82% D) 40.82%

Solve the problem.

13) A bank's loan officer rates applicants for credit. The ratings are normally distributed with a mean of 200 and a standard deviation of 50. If 40 different applicants are randomly selected, find the probability that their mean is above 215. 13) ________

A) 0.1179 B) 0.3821 C) 0.0287 D) 0.4713

Find the indicated probability.

14) Based on meteorological records, the probability that it will snow in a certain town on January 1st is 0.269. Find the probability that in a given year it will not snow on January 1st in that town. 14) ________

A) 1.269 B) 0.731 C) 0.368 D) 3.717

15) The table below describes the smoking habits of a group of asthma sufferers. 15) _______

	Nonsmoker	Occasional smoker	Regular smoker	Heavy smoker	Total
Men	356	42	70	44	512
Women	315	50	66	45	476
Total	671	92	136	89	988

If one of the 988 people is randomly selected, find the probability that the person is a man or a heavy smoker.

A) 0.564 B) 0.519 C) 0.494 D) 0.608

Determine whether the given procedure results in a binomial distribution. If not, state the reason why.

16) Rolling a single "loaded" die 40 times, keeping track of the numbers that are rolled. 16) _______

A) Not binomial: the trials are not independent.
B) Not binomial: there are more than two outcomes for each trial.
C) Procedure results in a binomial distribution.
D) Not binomial: there are too many trials.

Formulate the indicated conclusion in nontechnical terms. Be sure to address the original claim.

17) A psychologist claims that more than 47 percent of the population suffers from professional problems due to extreme shyness. Assuming that a hypothesis test of the claim has been conducted and that the conclusion is failure to reject the null hypothesis, state the conclusion in nontechnical terms. 17) _______

A) There is not sufficient evidence to support the claim that the true proportion is less than 47 percent.
B) There is sufficient evidence to support the claim that the true proportion is less than 47 percent.
C) There is not sufficient evidence to support the claim that the true proportion is greater than 47 percent.
D) There is sufficient evidence to support the claim that the true proportion is greater than 47 percent.

Assume that a hypothesis test of the given claim will be conducted. Identify the type I or type II error for the test.

18) A researcher claims that the amounts of acetaminophen in a certain brand of cold tablets have a standard deviation different from the $\sigma = 3.3$ mg claimed by the manufacturer. Identify the type II error for the test. 18) _______

A) The error of failing to reject the claim that the standard deviation is 3.3 mg when it is actually different from 3.3 mg.
B) The error of rejecting the claim that the standard deviation is more than 3.3 mg when it really is more than 3.3 mg.
C) The error of rejecting the claim that the standard deviation is 3.3 mg when it really is 3.3 mg.

1) B
2) C
3) D
4) C
5) B
6) B
7) B
8) B
9) B
10) D
11) C
12) A
13) C
14) B
15) A
16) B
17) C
18) A

Introductory Statistics
Quick Reference & R Commands

by Anthony Tanbakuchi. Version 1.1
http://www.tanbakuchi.com
ANTHONY@TANBAKUCHI·COM

Get R at: http://www.r-project.org
More R help & examples at:
http://tanbakuchi.com/Resources/R_Statistics/RBasics.html
R commands: **bold text**

1 Misc R

To make a vector / store data: **x=c(x1, x2, ...)**
Get help on function: **?functionName**
Get column of data from table: **tableName$columnName**
List all variables: **ls()**
Delete all variables: **rm(list=ls())**

$$\sqrt{x} = \texttt{sqrt(x)} \tag{1}$$
$$x^n = \texttt{x\^{}n} \tag{2}$$
$$n = \texttt{length(x)} \tag{3}$$
$$T = \texttt{table(x)} \tag{4}$$

2 Descriptive Statistics

2.1 Numerical

Let **x=c(x1, x2, x3, ...)**

$$\text{total} = \sum_{i=1}^{n} x_i = \texttt{sum(x)} \tag{5}$$
$$\min = \texttt{min(x)} \tag{6}$$
$$\max = \texttt{max(x)} \tag{7}$$
$$\text{six number summary}: \texttt{summary(x)} \tag{8}$$
$$\mu = \frac{\sum x_i}{N} = \texttt{mean(x)} \tag{9}$$
$$\bar{x} = \frac{\sum x_i}{n} = \texttt{mean(x)} \tag{10}$$
$$\tilde{x} = P_{50} = \texttt{median(x)} \tag{11}$$
$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} \tag{12}$$
$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \texttt{sd(x)} \tag{13}$$
$$CV = \frac{\sigma}{\mu} = \frac{s}{\bar{x}} \tag{14}$$

2.2 Relative standing

$$z = \frac{x-\mu}{\sigma} = \frac{x-\bar{x}}{s} \tag{15}$$

Percentiles

$$P_k = x_i, \quad \text{(sorted } x)$$
$$k = \frac{i-0.5}{n} \cdot 100\% \tag{16}$$

To find x_i given P_k, i is:

$$1: L = \frac{k}{100\%} \cdot n \tag{17}$$

2 : if L is an integer: $i = L + 0.5$; otherwise i=L and round up.

2.3 Visual

All plots have optional arguments:
main="" sets title
xlab="", ylab="" sets x/y-axis label
type="p" for **p**oint plot
type="l" for **l**ine plot
type="b" for **b**oth points and lines
Ex: `plot(x, y, type="b", main="My Plot")`
Histogram: **hist(x)**
Stem & leaf: **stem(x)**
Box plot: **boxplot(x)**
Barplot: **plot(T)** (where `T=table(x)`)
Scatter plot: **plot(x,y)** (where `x, y` are ordered vectors)
Time series plot: **plot(t,y)** (where `t, y` are ordered vectors)
Graph function: **curve(expr, xmin, xmax)** plot `expr` involving x

2.4 Assessing Normality

Q-Q plot: **qqnorm(x); qqline(x)**

3 Probability

Number of successes x with n possible outcomes.
(Don't double count!)

$$P(A) = \frac{x_A}{n} \tag{18}$$
$$P(\bar{A}) = 1 - P(A) \tag{19}$$
$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \tag{20}$$
$$P(A \text{ or } B) = P(A) + P(B) \qquad \text{if } A, B \text{ mutually exclusive} \tag{21}$$
$$P(A \text{ and } B) = P(A) \cdot P(B|A) \tag{22}$$
$$P(A \text{ and } B) = P(A) \cdot P(B) \qquad \text{if } A, B \text{ independent} \tag{23}$$
$$n! = n(n-1)(n-2)\cdots 2 \cdot 1 = \texttt{factorial(n)} \tag{24}$$
$${}_nP_k = \frac{n!}{(n-k)!} \qquad \text{Perm. no elements alike} \tag{25}$$
$$= \frac{n!}{n_1! n_2! \cdots n_k!} \qquad \text{Perm. } n_1 \text{ alike, } \ldots \tag{26}$$
$${}_nC_k = \frac{n!}{(n-k)!k!} = \texttt{choose(n,k)} \tag{27}$$

4 Random Variables

4.1 Discrete distributions

$$P(x_i): \text{ probability distribution} \tag{28}$$
$$E = \mu = \sum x_i \cdot P(x_i) \tag{29}$$
$$\sigma = \sqrt{\sum (x_i - \mu)^2 \cdot P(x_i)} \tag{30}$$

4.2 Continuous distributions

CDF $F(x)$ gives area to the left of x, $F^{-1}(p)$ expects p is area to the left.

$$f(x): \text{ probability density} \tag{31}$$
$$E = \mu = \int_{-\infty}^{\infty} x \cdot f(x)\, dx \tag{32}$$
$$\sigma = \sqrt{\int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\, dx} \tag{33}$$
$$F(x): \text{ cumulative prob. density (CDF)} \tag{34}$$
$$F^{-1}(x): \text{ inv. cumulative prob. density} \tag{35}$$
$$F(x') = \int_{-\infty}^{x'} f(x)\, dx \tag{36}$$
$$p = P(x < x') = F(x') \tag{37}$$
$$x' = F^{-1}(p) \tag{38}$$
$$p = P(x > a) = 1 - F(a) \tag{39}$$
$$p = P(a < x < b) = F(b) - F(a) \tag{40}$$

4.3 Sampling distributions

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{41}$$
$$\mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \tag{42}$$

4.4 Binomial distribution

$$\mu = n \cdot p \tag{43}$$
$$\sigma = \sqrt{n \cdot p \cdot q} \tag{44}$$
$$P(x) = {}_nC_x p^x q^{(n-x)} = \texttt{dbinom(x, n, p)} \tag{45}$$

4.5 Poisson distribution

$$P(x) = \frac{\mu^x \cdot e^{-\mu}}{x!} = \texttt{dpois(x, }\mu\texttt{)} \tag{46}$$

4.6 Normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \tag{47}$$
$$p = P(z < z') = F(z') = \texttt{pnorm(z')} \tag{48}$$
$$z' = F^{-1}(p) = \texttt{qnorm(p)} \tag{49}$$
$$p = P(x < x') = F(x') = \texttt{pnorm(x', mean=}\mu\texttt{, sd=}\sigma\texttt{)} \tag{50}$$
$$x' = F^{-1}(p) = \texttt{qnorm(p, mean=}\mu\texttt{, sd=}\sigma\texttt{)} \tag{51}$$

4.7 t-DISTRIBUTION

$$p = P(t < t') = F(t') = \texttt{pt(t', df)} \tag{52}$$
$$t' = F^{-1}(p) = \texttt{qt(p, df)} \tag{53}$$

4.8 χ^2-DISTRIBUTION

$$p = P(\chi^2 < \chi^{2'}) = F(\chi^{2'}) = \texttt{pchisq(}X^{2'}\texttt{, df)} \tag{54}$$
$$\chi^{2'} = F^{-1}(p) = \texttt{qchisq(p, df)} \tag{55}$$

4.9 F-DISTRIBUTION

$$p = P(F < F') = F(F') = \texttt{pf(F', df1, df2)} \tag{56}$$
$$F' = F^{-1}(p) = \texttt{qf(p, df1, df2)} \tag{57}$$

5 Estimation

5.1 CONFIDENCE INTERVALS

$$\text{proportion: } \hat{p} \pm E, \qquad E = z_{\alpha/2} \cdot \sigma_{\hat{p}} \tag{58}$$
$$\text{mean (σ known): } \bar{x} \pm E, \qquad E = z_{\alpha/2} \cdot \sigma_{\bar{x}} \tag{59}$$
$$\text{mean (σ unknown, use s): } \bar{x} \pm E, \qquad E = t_{\alpha/2} \cdot \sigma_{\bar{x}}, \qquad df = n-1 \tag{60}$$
$$\text{variance: } \frac{(n-1)s^2}{\chi^2_R} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_L}, \qquad df = n-1 \tag{61}$$

5.2 CRITICAL VALUES

$$z_{\alpha/2} = P(z > \alpha) = \texttt{qnorm(1-alpha/2)} \tag{62}$$
$$t_{\alpha/2} = P(t > \alpha) = \texttt{qt(1-alpha/2, df)} \tag{63}$$
$$\chi^2_L = P(\chi^2 < \alpha) = \texttt{qchisq(alpha/2, df)} \tag{64}$$
$$\chi^2_R = P(\chi^2 > \alpha) = \texttt{qchisq(1-alpha/2, df)} \tag{65}$$

5.3 REQUIRED SAMPLE SIZE

$$\text{proportion: } n = \hat{p}\hat{q}\left(\frac{z_{\alpha/2}}{E}\right)^2 \quad (\hat{p} = \hat{q} = 0.5 \text{ if unknown}) \tag{66}$$
$$\text{mean: } n = \left(\frac{z_{\alpha/2} \cdot \hat{\sigma}}{E}\right)^2 \tag{67}$$

6 Hypothesis Tests

`alternative` can be:
`"two.sided"`, **`"less"`**, **`"greater"`**
Test statistic and R function (when available) are listed for each.

6.1 1-SAMPLE PROPORTION

$H_0 : p = p_0$
`prop.test(x, n, p=`p_0`, alternative="two.sided")`

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} \tag{68}$$

6.2 1-SAMPLE MEAN (σ KNOWN)

$H_0 : \mu = \mu_0$

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{69}$$

6.3 1-SAMPLE MEAN (σ UNKNOWN)

$H_0 : \mu = \mu_0$
`t.test(x, mu=`μ_0`, alternative="two.sided")`
Where **`x`** is a vector of sample data.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \qquad df = n-1 \tag{70}$$