WEEK 6: CHAPTER 6, HYPOTHESIS TESTING

TABLE OF CONTENTS

Page:

List of Figures

List of Tables

List of m.files

Assignment

Required reading

! Larsen, R. J. and M. L. Marx. 2006. An introduction to mathematical statistics and its applications, 4th edition. Prentice Hall, Upper Saddle River, NJ. 920 pp.

" Read chapter 6, but skip p. 455-466 (decision rules for nonnormal data & the generalized likelihood ratio)

Understanding by Design Templates

Understanding By Design Stage 1 — Desired Results Week 6

LM Chapter 6 Hypothesis Testing Read 6.1-6.4, 6.6 {Skip 6.5}

G Established Goals

- Learn the standard hypothesis testing procedure with decision rules, null & alternative hypotheses, critical values, confidence limits, Type I and Type II error and power.

U Understand

- Failure to reject the null hypothesis does not imply that the null hypothesis is true.
- Factors controlling the power of a test.

Q Essential Questions

- Why do so many statisticians object to Neyman-Pearson hypothesis tests, and what is the alternative?
- What are null and alternative hypotheses?
- The *p-value* is a probability, but of what?
- If we reject the null hypothesis with an α-level of 5%, is the probability that the alternative hypothesis is true ≥ 95%?
- How can the power of a test be improved?

K *Students will know how to define (in words or equations)*

- **alpha level**, **alternative hypothesis**, **critical region**, **critical value**, **decision rule**, **hypothesis testing**, **null hypothesis**, **one-sided** vs. **two-sided alternative hypotheses**, ***p-value***, **positive predictive value**, **significance level**, **statistically significant**, **test statistic**, **Type I** & **Type II error**, **Z Ratio**

S *Students will be able to*

- Carry out and interpret hypothesis tests using the one-sample *z* test, and the one-sample binomial test
- Understand and calculate the probability of Type II error and power of a test
- Create one-sided and two-sided power curves for a test and understand the factors controlling the relative power of a test

Understanding by Design Stage 2 — Assessment Evidence Week 6 (7/5-7/11 M)
Chapter 6

- **Post in the discussion section by 7/13 W 10 PM** by W
  - Read **Sterne & Smith's (2001)** "What's wrong with significance tests?" and post an answer to that question in the weekly discussion board.
- **HW 6 Problems due Wednesday 7/13/11 W 10 PM**
  - **Basic problems (4 problems 10 points)**
    - **Problem 6.2.4**. Write your own program or just use Matlab's ztest.m; see how it is called in LMex060202_4th.m
    - **Problem** 6.3.2 Pawedness. Use LMcs060301_4th.m as a model
    - **Problem** 6.4.6 part c only, use LMEx060401_4th as a model.
    - **Problem 6.4.8** use LMEx060401_4th as a model, convert from power to Type II error
  - **Advanced problems (2.5 points each)**
    - **Problem 6.4.4** Use LMFig060405_4th.m as a model.
    - **Problem 6.4.18** Suggest using LMex060403_4th.m as a model
  - **Master problems (1 only, 5 points)**
    - Solve problem 6.2.2 include a graph that shows the two critical regions. Modify LMFig060203a_4th.m for the graph

Introduction

http://ts1.mm.bing.net/images/thumbnail.aspx?q=873051655724&id=0806c2adef25ee6e127138fb01

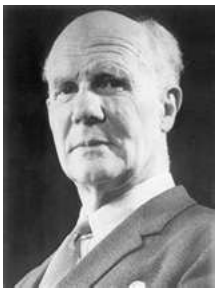

Figure 2. Egon Pearson

Figure 1. Jerzy Neyman

http://www.ds.unifi.it/gmm/statmodr/_images/jerzy.png

Chapter 6 is a quick, relatively non-mathematical (no calculus) introduction to hypothesis testing. It introduces the one-sample binomial and *z* test and introduces the concepts of null and alternative hypotheses, critical values, Type I and Type II error and power. For the opening pictures of statisticians, this chapter should have had pictures of the Polish and English statisticians Jerzy Neyman and Egon Pearson because they are are chiefly responsible for our current practice hypothesis testing. **Dennis (1996)** in a rousing defense of hypothesis testing against an attack by Bayesians refers to the paradigm presented in Chapter 6 as Fisher-Neyman-Pearson-Wald (-Rao -Efron) frequentist statistics. **Mayo (1996)** praises the work of Neyman & Pearson, especially Pearson, in proposing their hypothesis testing method and their invention of confidence intervals. Confidence intervals, which Larsen & Marx introduced in Chapter 5, were first described by Neyman in a 1934 talk, but the concept like hypothesis testing in general was controversial from the beginning. **Salsburg (2001, Chapter 12)** describes Pearson's first presentation of confidence limits only to be accused of attempting to fool the audience with a 'confidence trick.' Fisher never accepted Neyman & Pearson's approach to hypothesis testing and confidence limits. Neyman & Pearson would enjoy chapter 6, but Fisher might object to the sections on confidence intervals.

Many modern statisticians have rejected some features of hypothesis testing especially the decision rule, whereby results are divided into 'significant' and 'non-significant.' A Google

search under 'What's wrong with hypothesis testing?' brings up hundreds of web pages, but one of the best is the link to the paper by **Sterne & Smith (2001)**, who advocate abandoning the adjective significant from the statistician's lexicon. Nevertheless, most practicing environmental scientists follow the Neyman-Pearson frequentist theory of hypothesis testing. Many times a question from the audience at a meeting is, "Is that result significant?" By that, the questioner is asking whether the speaker had followed the standard decision rule and had been able to reject the appropriate null hypothesis at an $\alpha = 0.05$ level.

As noted by **Salsburg (2001)**, Neyman-Pearson hypothesis testing is the foundation for statistical analysis used by the FDA and the EPA. **Mayo (1996)** gives a rousing defense of the Neyman-Pearson hypothesis testing paradigm in her book, "Error and the Growth of Knowledge," which is a mixture of philosophy, history of statistics, prescription for how statistics and science should be done, and a polemic against Bayesian analysis. **Dennis (1996)** vigorously defends frequentist statistics as used by ecologists.

For this week's discussion, I want you to read and comment on **Sterne & Smith's (2001)** article, "What's wrong with significance testing?" Ecologists may want to read **Dennis (1996)** on "Should ecologists become Bayesians?," which also addresses the controversy over hypothesis testing. So, with that in mind, let's cover chapter 6 on hypothesis testing.

Definitions and Theorems

"The process of dichotomizing the possible conclusions of an experiment and then using the theory of probability to choose one option over the other is known as **hypothesis testing**. The two competing propositions are called the **null hypothesis** (written H_o) and the **alternative hypothesis** (written H_1 [or H_a])"

Comment (p. 433). If H_o: $\mu=\mu_o$ is rejected using a 0.05 **decision rule**, we say that the difference between $\bar{y}$ and μ_o is **statistically significant** ... **decision rules** are statements that spell out the conditions under which a null hypothesis is to be rejected."

Gallagher's comments on the decision rule: The **decision rule** should explicitly identify **IN ADVANCE** how the data are to be collected, the test to be used, the alternative hypothesis (*i.e.*, one-tailed left, one-tailed right or two-tailed), and the α-level for the test. Yes, in frequentist statistics, the *p value* depends on how you planned your experiment. **Berger & Berry (1988)** in a critique of frequentist statistics note that *p-values* depends on how the data were collected. They present the hypothetical case of a test of vitamin C vs. a Placebo in which 17 pairs of subjects were given placebos and Vitamin C. In 13 of the 17 trials, the subject taking Vitamin C improved to a greater extent, producing a p value of 0.049 (sum(binopdf([0:4 13:17],17,0.5))). If on the other hand, the investigators had designed the study such that they continued assigning to pairs Vitamin C & placebo until there were 4 successes in each group, then the study might produce exactly the same result 13 successes for Vitamin C and only 4 for the placebo. The *p values* would be different. Even though the data might be the same — 4 successes for Placebo and 13 for Vitamin C — the **p value** is now 0.021 (**Berger & Berry 1998, Figure 2**).

Much of statistical inference and the scientific method is based on the logical syllogism called the **modus tollens**:

modus tollens:	
If	**A**
Then	**B**
Observe	**Not B** (**~B**)
Conclude	**Not A**

Observing **not B** (often abbreviated ~B) allows us to reject **A**. Observing **B** allows us neither to accept nor reject **A**. **A** in the case of hypothesis testing is the null hypothesis including the underlying probability model for the data. For example **A** could comprise 4 components, using the gas mileage example from **Larsen & Marx (2006, Chapter 6)**:

① The null hypothesis H_o: $\mu_o = 25$ mpg, $\sigma = 2.4$
② The underlying probability model and proposed test. In this case, the data will be analyzed with a one-sample *z* test against a right-tailed alternative hypothesis $\mu > 25$ mpg.
③ the assumptions of the statistical test are met or don't matter. For the *z* test, the data must represent a random sample from a normally distributed population
④ an α level of 0.05 will be used to set the probability of Type I error.

Then, **B** is the predicted outcome of the statistical test if the null hypothesis specified in **A** is true and the other 3 conditions in **A** are satisfied. Now **A** and **B** must be set in advance before examining the data, or better yet, before the experiment or survey is even conducted. Once the data are collected and analyzed they will either meet the conditions in **B** or be classified as **not B**.

After the experiment or survey has been performed, the data tested for assumptions, and analyzed, the result will be **B** or **not B**, but the decision between **B** and **not B** is based on the p value, the probability of the observed test statistic or a test statistic more extreme if the null hypothesis is true. The *p value* is **NOT** the probability of the null hypothesis. **B** can take three equivalent forms: ❶ the test statistic will have a *p-value* $\geq \alpha$ level, ❷ the test statistic will be outside of the critical region when expressed in the natural scale or ❸ the test statistic, expressed as a z-ratio (or t ratio, or F ratio), lies outside of the critical region, e.g., $Z \geq z_\alpha$ (left-tailed), $Z \leq z_{1-\alpha}$ (right-tailed) or $z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}$ (two-tailed). If **B** is observed, one fails to reject the null hypothesis specified in **A** and reports the **p-value** of the test, "I failed to reject the null hypothesis at the 5% significance level (right-tailed *z* test, $\bar{X}$ =25.25, n=30, μ=25, σ=2.4, p=0.284)."

Not B,the complement of **B**, can take three equivalent forms: ❶ the test statistic will have a *p-value* < α level, ❷ the test statistic will be within the critical region when expressed in the natural scale or ❸ the test statistic, expressed as a *z*-ratio (or *t* ratio, or *F* ratio), lies within the critical region: e.g, $Z < z_\alpha$ (left-tailed), $Z > z_{1-\alpha}$ (right-tailed) or $Z < z_{\alpha/2}$ or $Z > z_{1-\alpha/2}$ (two-tailed). If

Not B is observed, the probability of observing the actual data or data more extreme is very unlikely if the null hypothesis is true. One rejects the null hypothesis specified in **A** at the α-level of significance, e.g., "I rejected the null hypothesis at the 5% significance level (right-tailed *z* test, $\bar{X}$ =26.8, n=30, μ=25, σ=2.4, p=0.0003)." Many scientists reduce the α level of the test after the fact to match the ***p-value***, concluding "I rejected the null hypothesis at the 0.1% significance level (1-tailed *z* test, p=0.0003)." This is unfortunate since the *p values* have their proper meaning only if the **decision rule** including the α level, test statistic, and alternative hypotheses is set in advance. You are not allowed to change the α-level to a higher or lower level or to change from a two-tailed to a one-tailed alternative hypothesis after you've analyzed the data. The **decision rule** requires that the α-level and alternative hypothesis be set before you analyze the data. You shouldn't report an α level of 0.001 as being the criterion for statistical significance if the *a priori* decision rule would have permitted rejecting the null with an alpha level of 0.05. Reporting the actual *p value* and sample size for the test is sufficient information for the reader of your work to evaluate the strength of evidence against the null hypothesis. One doesn't need to imply, "In my lab the tests have such power than only α levels of 0.001 are used.")

There is a third possibility as well. After collecting the data and analyzing them, it may turn out that the assumptions of the test, indicated in **A③** above have been grossly violated. For example, an analysis of the data may indicate such severe departures from normality that a z test can not be used. In that case, one stops the hypothesis test and creates a new decision rule based on transformed data or a new statistical test. In hypothesis testing, one must check the assumptions of the test and adjust the test statistic accordingly. One is not allowed to change the α level after the data have been analyzed.

Mayo (1996, Chapter 5) describes a slightly different model for performing hypothesis testing in the Neyman-Pearson framework. Based on an earlier analysis by **Suppes (1969)**, Mayo argues that there are three hierarchic and linked models that are evaluated in hypothesis testing: the **primary**, **experimental**, and **data** models. The **primary model** links the hypothesis with the larger area of inquiry and includes the specification of the null hypothesis. For example, the theory of evolution by means of natural selection was tested by evaluating whether a latitudinal cline in fruitfly body size evolved from a single population of fruitflies introduced in South America in 1978. **Huey et al. (2000)** tested three null hypotheses. First, they tested whether regression of body size versus latitude in North American (NA) flies had zero slope (i.e, the flies hadn't evolved in the two decades since their introduction). Second, they tested whether NA male and female flies evolved (or failed to evolve) at the same rate (*i.e.*, the slope of size versus latitude was the same for male and female flies), and third that NA flies evolved to have the same body size versus latitude regression as European (EU) flies. **Mayo's (1996)** & **Suppes' (1969)** **experimental model** includes the research procedure detailing how the data are to be collected and would specify choice of experimental model, sample size, experimental variables, test statistics, alternative hypotheses, and the α-level for evaluating the null hypothesis. In **Huey et al.'s (2000)** fruitfly study, flies were to be collected from different latitudes and reared for several generations under identical conditions in the lab to ensure that the body size differences weren't due to difference in food supply or temperature in different regions. Finally the **data model** puts the data in proper form for applying the analytical methods, including testing

whether the assumptions of statistical model are met. In **Huey et al.'s (2000)** study, the fruitfly wing lengths had to be log-transformed to meet the linearity and equal variance assumptions of ordinary least squares regression (to be covered in **Larsen & Marx's (2006) Chapter 11**).

In **Huey et al. (2000)**, the latitudinal gradient in fly-wing lengths revealed one of the most rapid rates of evolution ever recorded. Figure 3 shows the ln (wing size) versus latitude regression for North American and European fruitflies. The NA female slope (0.0020 ± 0.0004 {± standard error}, p<0.001) approximately equals the European female slope (0.0018 ± 0.0004, p<0.001). However, the decision rules of Neyman-Pearson statistical inference only allow **Huey et al. (2000)** to conclude that they **failed to reject the null hypothesis** that the slopes of North American and European female wing lengths were the same. The *modus tollens* provides the logical foundation for rejecting null hypotheses, but it doesn't provide logical justification for accepting null hypotheses. If **A** then **B**, observe **B**, then accept **A** is true is not a valid syllogism. Based on these data, we **can** conclude that the slopes of wing length versus latitude for both NA male and female flies are not zero. We can reject the null hypothesis that NA flies have not evolved. **Huey et al. (2000)** rejected the null hypothesis that NA male and female flies were evolving at the same rate because the wing-length slope vs. latitude for NA male flies (0.0007±0.0004, p=0.0265) was less than the slope for females (0.0020±0.0004). Science, through the *modus tollens*, advances through the rejection of false null hypotheses, but sometimes failure to reject null hypotheses can also provide convincing support for theories. It is amazing that those NA females evolved to resemble the European females.

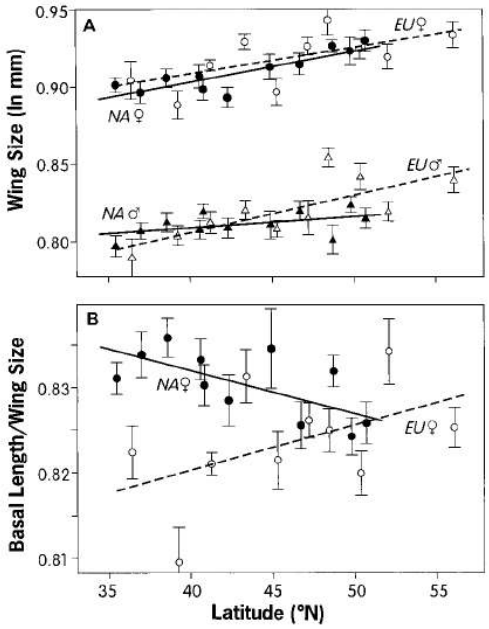

Figure 3. Top panel: Wing Size vs. latitude for North American and European flies from **Huey et al. (2000)**. Hypothesis tests led to the rejection of zero slope for North American female and male flies. The authors also rejected the null hypothesis that North American males and females had the same slope. The authors failed to reject the null hypothesis that North American and European female flies had the same slope.

If the null hypothesis is rejected, it used to be customary to state that the alternative hypothesis was significant. However, statisticians and natural scientists were troubled over using the same adjective 'significant' or the phrase 'signficant at an α-level of 0.05,' to describe a result with *p-value* 0.049 or 0.00001. **Sterne & Smith (2001)** and **Ramsey & Schafer (2002)** provide sliding scales based on *p-value* to express the strength of evidence against the null hypothesis. A *p-value* of 0.05 is regarded as 'moderate' evidence against the null hypothesis and *p-values* less than 0.001 are regarded as convincing or strong evidence against the null hypothesis.

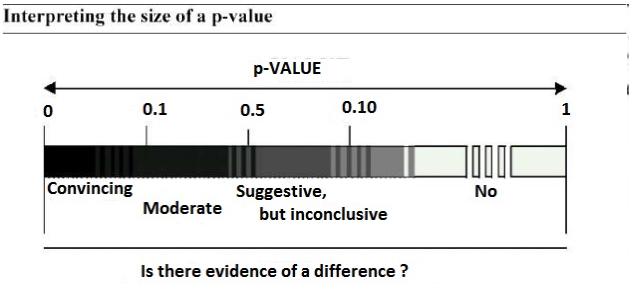

Figure 4. Adjectives to describe the strength of evidence against a null hypothesis from **Ramsey & Schafter (2002)**.

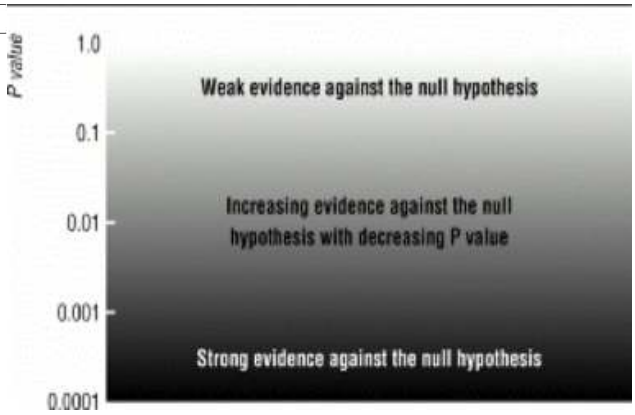

Figure 5. Adjectives to describe the strength of evidence against a null hypothesis from **Sterne & Smith (2001)**.

There are two troubling aspects about the use of the **Sterne & Smith (2001)** and **Ramsey & Schafer (2001)**. First, as noted in many letters to the British medical journal after the publication of **Sterne & Smith (2001)**, it is essential that the sample size be noted in reporting the results. This point is raised nicely by **Larsen & Marx (2006)** in Section 6.6. If the sample size is very large, even minor differences in a test statistic can have exceptionally low p-values. Second, there is a potential for weakening the strict guidance that the Neyman-Pearson decision rule provides. The scientific method depends on scientists being able to reject false hypotheses. It is the basis for Popper's scientific method based on Conjectures and Refutations. Figures 4 and 5 tend to weaken that key element of the scientific method, Popper's demarcation principle that tells a scientist the conditions under which he can confidently reject a proposition. In the above figures, a *p-value* of 0.05 provides just 'suggestive' evidence against the null hypothesis. The modern scientific method isn't based on a suggestion principle. Scientists don't perform experiments or surveys to provide mere suggestions that the null hypothesis is wrong; they choose an α level where a clear decision can be reached. Rather than using adjectives like 'suggestive' and 'moderate' to weaken the decision rule, perhaps more attention should be made in picking the α level for the test. If there is risk in having too high a Type II error, then recast the decision rule to increase the power of the test, for example by increasing the α level to 0.1 from 0.05.

Definition 6.2.1 Any function of the observed data whose numerical value dictates whether H_o is accepted or rejected is called a **test statistic**. The set of values for the test statistic that result in the null hypothesis being rejcted is called the **critical region** and is denoted *C*. The particular point in C that separates the rejection region from the acceptance region is called the **critical value**.

Definition 6.2.2 The probability that the test statistic lies in the critical region *when H_o is true* is called the **level of significance** and is denoted α. [significance level is also called **α-level**]

"If there is reason to believe before any data are collected that the parameters being tested is necessarily restricted to one particular "side" of H_o, then H_1 is defined tgo reflect that limitation

and we say that the alternative **hypothesis is one-sided**... If no such a priori information is available, the alternative hypothesis needs to accommodate the possibility that the true parameter value might lie one either side of μ_o. Any such alternative is said to be **two-sided**. For testing H_o: $\mu=\mu_o$, the two sided alternative is written H_1: $\mu \neq \mu_o$." (p 434)

Definition 6.2.3 The ***P-value*** associated with an observed test statistic is the probability of getting a value for that test statistic as extreme or more extreme than what was actually observed (relative to H_1) *given that H_o is true*. **Gallagher note:** one should **NOT** conclude that the **p-value** is the probability that the null hypothesis is true, nor should one conclude that the complement of the **p-value** (1 - **p-value**) is the probability that the alternative hypothesis is true. If one rejects the null hypothesis with a **p-value** of 0.04, ones should **NOT** conclude that there is a 96% chance that the alternative hypothesis is true. In fact, the probability may be no better than about 50% that the alternative hypothesis true. **Sterne & Smith (2001, Table 2, p. 228)** provide an example drawn from modern clinical medical practice in which they calculate the **positive predictive value** of a test, the probability that a significant result is indeed true. The **positive predictive value** is the complement of the probability of a false positive result (1-false positive probability). Even with an α-level of 0.05, the **positive predictive value** may only be about 50% if a field uses tests with low power (say 50%) and the initial probability of false null hypotheses is low (e.g, 10%). Under those conditions, only half the significant results published are true. **Ioannidis (2005)** pursues a similar argument to argue that the majority of published scientific research findings are false. So, don't be fooled into thinking that by rejecting a null hypothesis at the α-level of 0.05 that you have at least a 95% chance that your alternative hypothesis is true.

Type I and Type II error There are two kinds of errors that can be committed in the process of hypothesis testing. They are shown in the following table. **Type I error** is rejecting the null hypothesis H_o when H_o is true. Once a decision has been made to reject H_o, the probability of Type I error is the test statistic's ***p-value***. The symbol for the probability of committing **Type I error** is α. The probability of committing a Type I error is a test's **α-level** or **level of significance**. **Type II error** is the probability of failing to reject the null hypothesis H_o when H_o is false. The symbol for the probability of committing Type II error is β.

Table 1. Hypothesis testing decision tree, Type I and Type II errors. **Larsen & Marx (2006, p 447)**.

		True State of Nature	
		H_o is true	H_1 is true
Our Decision	Fail to reject H_o	Correct Decision	Type II error
	Reject H_o	Type I error	Correct Decision

If β is the probability that we fail to reject H_o when H_1 is true, then 1 - β is the probability of the complement, that we reject H_o when H_1 is true. We call 1 - β the **power** of the test; it represents the ability of the **decision rule** to "recognize" (correctly) that H_o is false.

Testing with the the one-sample Z test

Theorem 6.2.1 (**Larsen & Marx, 2006, p. 435**) introduces the 1-sample *Z* test. The *Z* arises because the *z* distribution is the conventional name for the standard normal distribution, with mean 0 and unit variance. The name may arise from the conventional labeling of the X-axis as 'z,' although **Larsen & Marx (2006)** usually label the abscissa with '*y*' and the ordinate as 'f(*y*)'

Theorem 6.2.1 Let $Y_1, Y_2, \ldots Y_n$ be a random sample of size n from a normal distribution where σ is known. Let $z = \dfrac{\bar{y} - \mu_o}{\sigma/\sqrt{n}}$.

- To test H_o: $\mu = \mu_o$ versus H_1: $\mu > \mu_o$ at the α level of significance, reject H_o if $z \geq z_\alpha$.
- To test H_o: $\mu = \mu_o$ versus H_1: $\mu < \mu_o$ at the α level of significance, reject H_o if $z \leq -z_\alpha$.
- To test H_o: $\mu = \mu_o$ versus H_1: $\mu \neq \mu_o$ at the α level of significance, reject H_o if *z* is either (1) $\leq -z_{\alpha/2}$ or (2) $\geq z_{\alpha/2}$

In most statistical testing, σ is not known and must be estimated from the sample or samples. In that case, the standard normal, or *z*, distribution is not appropriate. The usual distribution that is used to account for the reduced precision when σ must be estimated from a sample is Student's *t* distribution, invented by William S. Gossett of the Guinness Brewing Company, who published his famous *t* ratio under the pseudonym 'Student:' $t = \dfrac{\bar{y} - \mu_o}{s/\sqrt{n}}$. **Larsen & Marx (2006)** introduce the use of Student's t distribution with their Theorem 7.4.3.

Examples and Case Studies

Gas Fuel Additive Study

Throughout Chapter 6, Larsen & Marx use the example of the fuel additive study. Thirty cars were driven cross country to test a fuel additive designed to increase gas mileage. The gas mileage before the additive is μ = 25 with σ = 2.4. Assume σ is known to be 2.4. Then the probability density function is described by the normal curve f_Y:

$$f_Y(y;\ \mu) = \frac{1}{\sqrt{2\pi}(2.4)}\ e^{-\frac{1}{2}\left(\frac{y-\mu}{2.4}\right)^2}$$

If the existing gas mileage was 25 mpg, we are testing
H_o: $\mu = 25$
H_1: $\mu > 25$

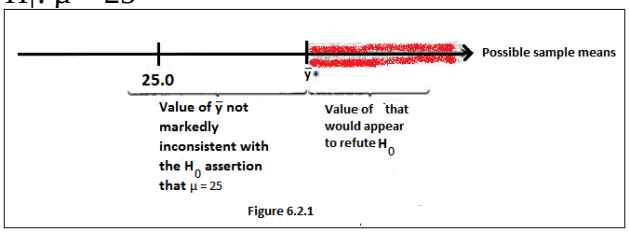

Figure 6.2.1

Is 25.25 a good choice for rejecting the null hypothesis? No. As shown in Figure 6, if the null hypothesis were true, then 28.43% of the samples of size 30 would have means ≥ 25.25.

Should the cutoff be 26.5%? As shown in Figure 7, only 0.03% of the area of the curve would have means ≥ 26.5 if the null hypothesis was true (μ=25; σ=24).

Figure 6.2.2

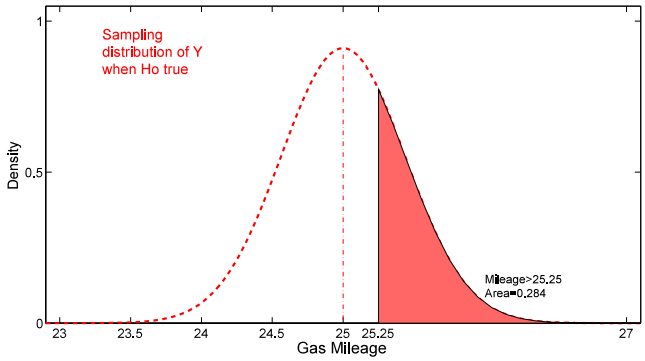

Figure 6. What percentage of means from a sample size 30 would have means ≥ 25.25 if the null hypothesis was true? 28.42%

Figure 6.2.3

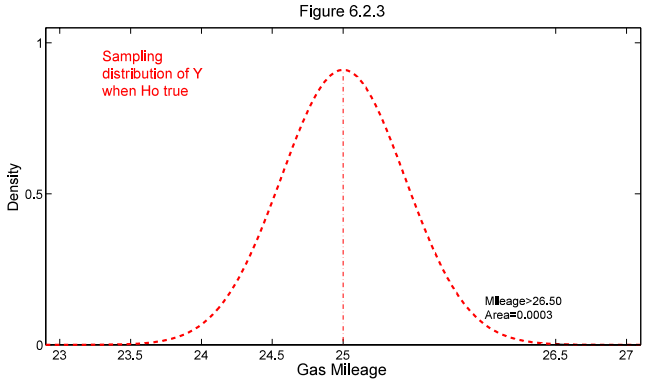

Figure 7. What percentage of means from a sample size 30 would have means ≥ 26.5 if the null hypothesis was true? .03%

What would the appropriate cutoff be? We can solve for the gas mileage such that 5% or fewer of the samples would have mileages greater than that critical value. That value is 25.7207 (norminv(0.95,25,2.4/sqrt(30)) and the relationship to the normal distribution is shown in Figure 8.

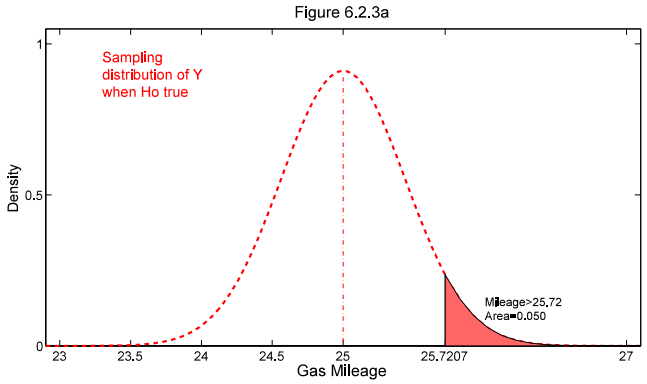

Figure 8. What percentage of means from a sample size 30 would have means ≥ 25.72 if the null hypothesis (μ=25, σ=2.4) was true? 5%

Z ratios & the gas-additive study

Decision rules and **critical values** can either be expressed in the natural scale, such as 25.7207 mpg or converted to z-ratios. A z-ratio is dimensionless since it is the ratio of a statistic, with appropriate units such as *mpg*, divided by the standard error for the statistic, which will always have the same units, such as *mpg*. In the case of the critical value for the fuel additive study, rejecting H_o: μ=25.0 mpg when

$$\bar{y} \geq \bar{y}^* = 25\ mpg + norminv(.95) \cdot \frac{2.4\ mpg}{\sqrt{30}} = 25\ mpg + 1.6449 \cdot \frac{2.4\ mpg}{\sqrt{30}} = 25.7207\ mpg$$

is clearly equivalent to rejecting H_o when $\dfrac{\bar{y} - 25\ mpg}{2.4\ mpg/\sqrt{30}} \geq 1.6449$.

With μ=25 & σ=2.4, the **decision rule** stated that H_o should be rejected at an α-level of 0.05 if $\bar{y}$ equaled or exceeded 25.7207 (25.781 due to rounding in the text. The probability of committing Type I error is set by the **decision rule** at 0.05:

$$\begin{aligned} &P(\text{Type I error}) = P(\text{reject } H_o \mid H_o \text{ is true}) \\ &= P(\bar{y} \geq 25.7207 \mid \mu = 25 \ \&\ \sigma = 2.4) \\ &= P\left(\frac{\bar{Y} - 25\ mpg}{2.4\ mpg/\sqrt{30}} \geq \frac{25.7207\ mpg - 25\ mpg}{2.4\ mpg/\sqrt{30}} \right) \\ &P(Z \geq 1.64) = 0.05 \end{aligned}$$

For example, what is the probability of committing a Type II error in the gasoline experiment if μ_o were 25 mpg, but the true μ (with the additive) were 25.750. By definition,

$$\begin{aligned} &P(\text{Type II error} \mid \mu = 25.750) = P(\text{fail to reject } H_o \mid \mu = 25.750 \ \&\ \sigma = 2.4) \\ &= P(\bar{Y} < 25.781 \mid \mu = 25.750 \ \&\ \sigma = 2.4) \\ &= P\left(\frac{\bar{Y} - 25.75}{2.4/\sqrt{30}} < \frac{25.718 - 25.75}{2.4/\sqrt{30}} \right) \\ &P(Z < -0.07) = 0.4721 \end{aligned}$$

So, even if the fuel additive increased gas mileage to 25.75 mpg, our **decision rule** would be tricked 47.21% of the time, telling us not to reject H_o. The symbol for the probability of committing type II error is β.

Figure 9 shows the sampling distribution of $\bar{y}$ when μ = 25 & σ = 2.4 and when μ = 25.75 & σ = 2.4 (H_1 is true).

Figure 10 shows the probability of Type II error if μ = 26.8 mpg.

"If β is the probability that we fail to reject H_o when H_1 is true, then 1 - β is the probability of the complement, that we reject H_o when H_1 is true. We call 1 - β the **power** of the test; it represents the ability of the **decision rule** to 'recognize' (correctly) that H_o is false." A **power curve** is a graph of 1 - β versus the set of all possible parameter values.

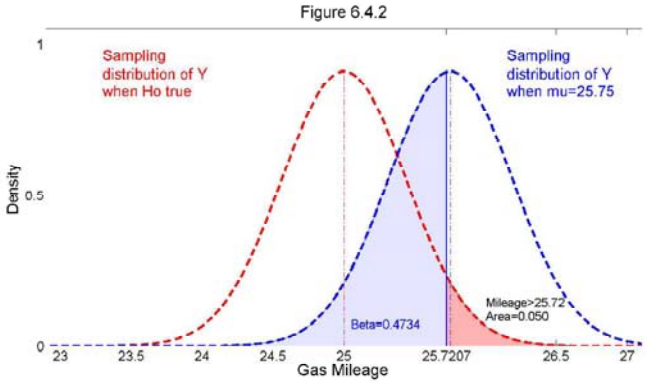

Figure 9. If μ=25.75, a decision rule based on H_o: μ=25.0, σ=2.4 & n=30 (i.e., critical value =25.2707= norminv(.95,25,2.4/sqrt(30))) would have a P (Type II error) = β = 0.4734 = normcdf(norminv(.95,25,2.4/sqrt(30)),25.75,2.4/sqrt(30))

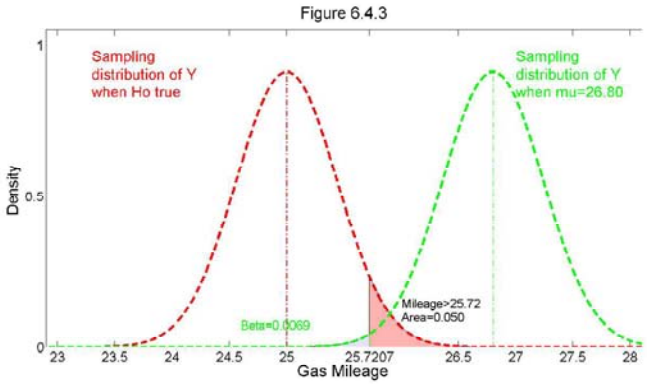

Figure 10. If μ=26.8, a decision rule based on H_o: μ=25.0, σ=2.4 & n=30 (i.e., critical value =25.2707= norminv(.95,25,2.4/sqrt(30))) would have a P (Type II error) = β = 0.0069 = normcdf(norminv(.95,25,2.4/sqrt(30)),26.8,2.4/sqrt(30))

Figure 11 shows the power curve for testing H_o: μ = 25 mpg & σ = 2.4 mpg vs. H_1: μ > 25 mpg & σ = 2.4 mpg.

Figure 12 shows the power curves for the gas mileage experiment comparing a 60-car and 30-car experiment. The power curves are shown for a 2-tailed test.

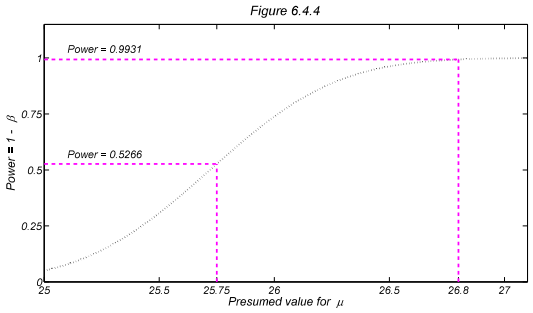

Figure 11. Power curve for H_o = 25, σ = 2.4 & n=30. Also shown are the estimated power of the test against alternative hypotheses μ = 25.75 and μ = 26.8

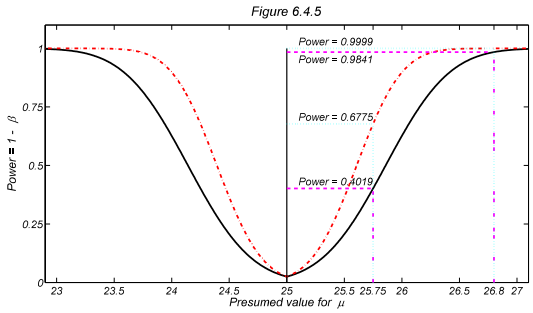

Figure 12. Two-tailed power curves for H_o = 25, σ = 2.4 & n=30 (black) and n=60 (red dashed lines). Also shown are the estimated power of each test against alternative hypotheses μ = 25.75 and μ = 26.8. With n=30, the power of the test against H_1: μ=25.75 is 0.4019, but with n=60, the power increases to 0.6775. The power also increases versus H_1: μ=26.8 from 0.9841 to 0.9989. The previous figure was the power curve for the 1-tailed test, which is more powerful than the 2-tailed test for equal sample size.

Figure 13 shows what happens to 1-β (when μ = 25.75) if σ, n, and μ are held constant but α is increased to 0.10 instead of 0.05.

Using the gasoline additive example, Figure 14 reveals the strong effects of reducing σ on the power of a test.

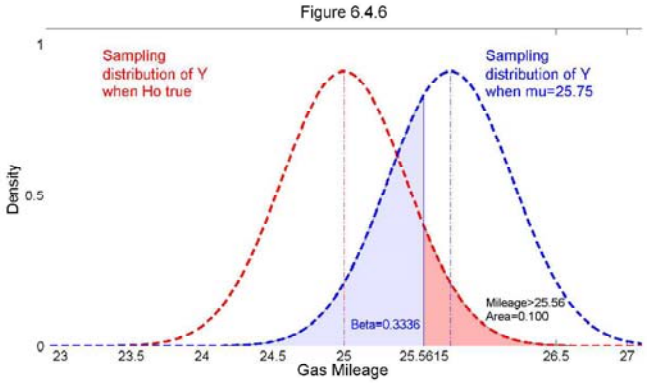

Figure 13. Increasing α from 0.05 to 0.10 decreases β from 0.4734 (see Figure 6.4.2) to 0.336, thus increasing power from 53% to 67%.

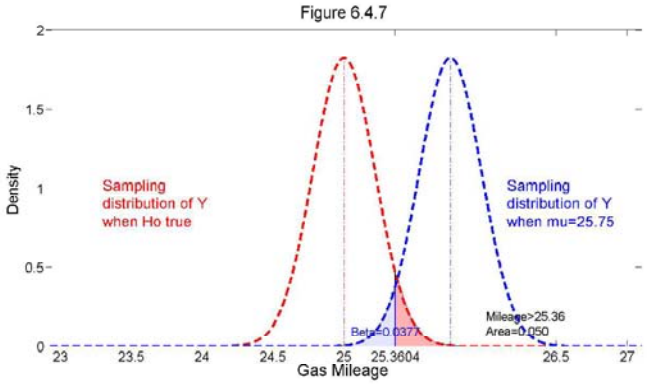

Figure 14. Decreasing σ from 2.4 to 1.2 decreases β from 0.4734 (see Figure 6.4.2) to 0.0377, thus increasing power from 53% to 96%.

Figure 15 shows the 1-tailed power curve for the gasoline additive data.

Figure 16 shows the 1-tailed power curves for the gasoline additive data for sample sizes of 30, 60 and 900.

Examples 6.2.1 & 6.2.2

Eighty six students with typical scores from a high school were taught with a new curriculum. Their SAT-1 math scores at the end of the instructional period was 502. Given that the national average on the SAT-1 math score is 494 with σ = 124, are the scores that these students earned greater or less than expectation? The null hypothesis is that $\mu = \mu_o = 494$.

The phrasing of the question implies a two-tailed alternative test. This can be tested at the α=0.05 level of significance by computing the z score and comparing it to the critical value for a two-tailed z test of 1.96. If the z ratio is less than -1.96 or greater than 1.96, the conclusion would be to reject the null hypothesis.

The z ratio for these data is

$$z = \frac{502 - 494}{124/\sqrt{86}} = 0.60$$

, which is within the two critical values of -1.96 and 1.96, so the conclusion is fail to reject. Figure 17 shows the z-ratio=0.60 relative to the two critical regions.

What is the p-value of the test? The *p-value* is defined above and is the smallest α level at which these data can be used to reject H_o. As programmed in LMex060202_4th.m, the two-

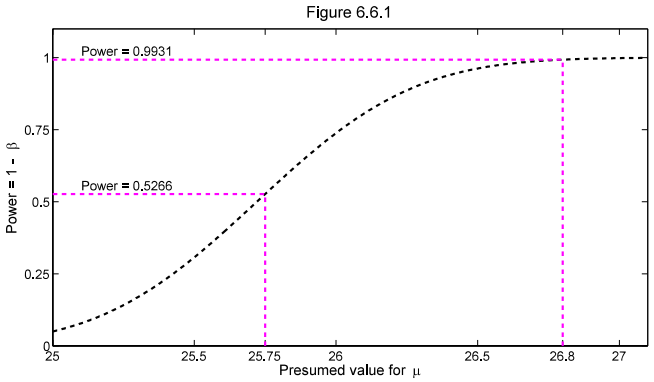

Figure 15. One-tailed power curve for H_o = 25, σ = 2.4 & n=30. Also shown is the estimated power of the test against alternative hypotheses μ = 25.75 and μ = 26.8. With n=30, the power of the 1-tailed test against H_1: μ=25.75 is 0.5266. The two-tailed test shown in a previous figure had a power of only 0.4019. The power of the 1 tailed test against H_1: μ=26.8 is 99.31%, which is an increase from 98.41%. Tests against 1-tailed alternatives are more powerful than two-tailed tests.

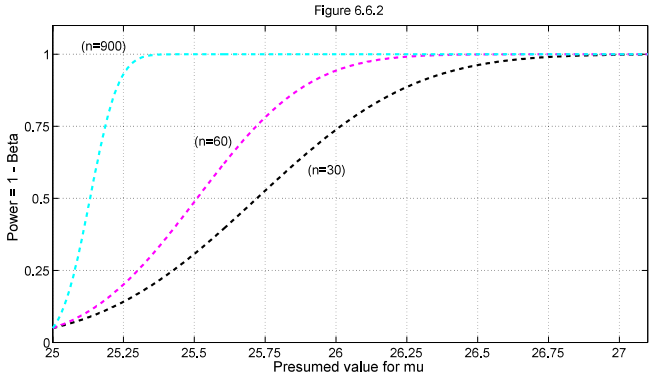

Figure 16. One-tailed power curves for H_o = 25, σ = 2.4 & n=30, 60 and 900.

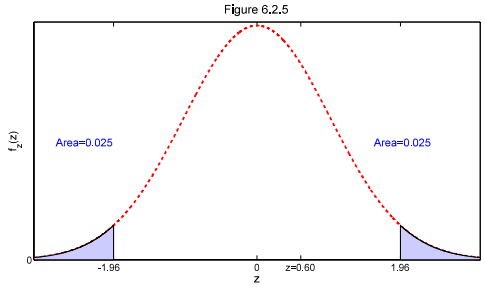

Figure 17. The z ratio of 0.6 is within the two shaded critical regions, so the decision is 'fail to reject H_o'

tailed p-value associated with the z ratio = 0.6 can be found directly from the Matlab's one-sample z test (ztest.m):

```
[H,P,CI,ZVAL] = ztest(repmat(Ybar,n,1),muo,sigmao,alpha,'both')
fprintf('The two-tailed p-value is %6.4f\n',P);
```

Or, the 1-tailed p-value can be found from the cumulative normal distribution function and multiplied by two. Note that the program must check whether z is less than 0 before determining whether the normal cumulative distribution function should be used to find the left tail or the right tail of the distribution:

```
if z<0
   pvalue=normcdf(z);
else
   pvalue=1-normcdf(z);
end
fprintf('The two-sided p-value is %6.4f\n',pvalue*2)
```

Figure 18 shows that the above statement finds the upper portion of the normal cumulative distribution function.

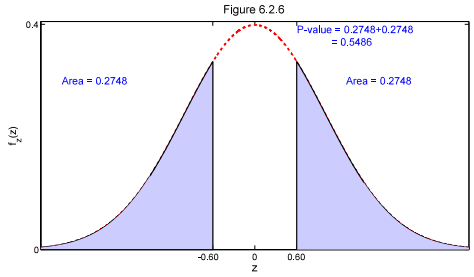

Figure 18. The p-value is the region to the right of abs(z ratio) and to the left of -abs (z ratio). The z ratio was 0.6 producing a two-tailed *p-value* of 0.5486.

Case Study 6.3.1 Point Spread

Point spreads are designed so that the both teams should have an equal probability of winning once the point spread has been added to the team's points. In a study of 124 NFL games, the favored team beat the spread in 67 games (54%) of the time. Is this more than could be expected by chance? Apply the one sample binomial test, two tailed with an α level of 0.05.

I've written a program called onesamplebinom.m which solves the one-sample binomial test with both the approximate solution, as described in the text (4th edition page 441-442), and using the exact test based on the binomial pdf. As described in the text, there is a test to determine whether the normal approximation is adequate. Here is the call to binomonesmple:

```
X=67; n=124;p=0.5;
[p2tailed,exact,zval]=onesamplebinom(n,X,p,usebinomial);
```

```
Passed the test for the normal approximation
Using the normal approximation for the 1-sample binomial, z=0.8980 and 2-tailed p=0.3692
Using the binomial pdf for the 1-sample binomial, 2-tailed p=0.4191
Using the binomial pdf for the 1-sample binomial, 1-tailed p=0.2095
```

Using the **decision rule** and an α-level of 0.05, I conclude that there is little evidence to reject the null hypothesis (one-sample binomial test, z = 0.89, approximate two-tailed *p-value* = 0.37). Note that if I were doing this analysis for real, I would use the exact two-tailed test, reporting again that there is little evidence to reject the null hypothesis (one sample binomial test, exact two-tailed *p-value* = 0.42).

Brown et al. (2001) have shown that the standard methods for generating 95% confidence limits using the large sample approximation are invalid. **Larsen & Marx (2006)** don't generate confidence limits for their approximation.

Case Study 6.3.2 Deaths after birthdays

Do people hang on until after their birthdays to die? Among 747 decedents reported in a Salt Lake newspaper, only sixty (8%) had died in the three months prior to their birthday. The expected value would be 25%. Could this result have been due to chance?

I can solve that with binomonesample, implemented below in LMcs060302_4th.m
X=60; n=747; p=3/12; [p2tailed,exact,zval]=onesamplebinom(n,X,p,usebinomial);
which provides the following output
DeMoivre-Laplace rule is met, Normal approximation probably ok
Passed the test for the normal approximation
Normal approximation for the 1-sample binomial test, z=-10.7099, 1-tailed p= 4.6e-027
Using the binomial pdf, Lower Tail p= 3.9e-033

Based on these results, I'd conclude that there was overwhelming evidence to reject the null hypothesis (one-sample binomial test, 1-tailed $p < 10^{-6}$). In general, it isn't a good idea to report p-values as extreme as 5×10^{-27}. With the exact test, while it would be warranted reporting such low p-values, $p < 10^{-6}$ should suffice as providing sufficient grounds for rejecting the null hypothesis.

Example 6.3.1

A drug is given to 19 elderly patients. The standard drug is effective in relieving arthritis pain in 85% of the cases. The researcher wishes to test the null hypothesis p=0.85 vs. the two-sided alternative hypothesis H_1: $p \neq 0.85$. The **decision rule** will be based on an exact one-sample binomial test. What are the critical values if the α level is to be approximately 10%?

I wrote a Matlab program (**LMex060301_4th.m**), that solves the problem. The lower critical value is 13 with a cumulative pdf of 0.0537. The upper critical value is 19 with cumulative pdf of 0.0456. With these critical values & H_o: p=0.85, P(Type I error)=alpha=0.0993. We can attempt to use this program to find the critical values for α = 0.05. The program again finds that the lower critical value is 13, but now it returns that there is no upper critical value. For α=0.20, the program finds that the upper critical value is 19 and that the lower critical value is 14. What is the exact 2-sided p-value if 19 of the elderly patients had responded favorably. It is 0.0993, calculated using: p2sided=sum(binopdf([0:13 19],19,0.85)
Why did I find the sum of the binomial pdf from 0 to 13? The expected number of successes is 0.85*19=16.15. If 19 were observed, that would be 2.85 patients more than expected. A two-tailed probability, by definition, tallies all of those events that are equal to or more extreme given the null hypothesis. We must consider those extreme cases on the other side of the null hypotheses. If 0 to 13 patients had responded favorably, the difference from expectation would

exceed 2.85, so the sum of the binomial probability for these variables comprise the lower tail of the two-sided p-value.

Example 6.4.1

With a null hypothesis of μ = 100 and σ = 14, how many samples are required at the α = 0.05 level of significance to achieve a power (1 - β) of 60% for the alternative hypothesis μ = 110. **Larsen & Marx (2006, p 454-455)** derive the answer, but it is solved in a few lines in Matlab:

```
muo=100;sigma=14;Power=.6;muh1=103;
N = sampsizepwr('z',[muo sigma],muh1,Power,[],'tail','right');
```

The answer in Matlab is 79, but in Larsen & Marx is 78. Carrying out the Larsen & Marx calculations to full double precision reveals that N=78.1925 samples are required to achieve a power equal to 60%, but since fractional samples aren't allowed, 79 is the proper answer. The same Matlab function can be used to calculate the power for 78 samples:

```
n=78;Power = sampsizepwr('z',[muo sigma],muh1,[],n,'tail','right');
```

The answer returned by Matlab is 59.78%. The sampsizepwr function can also be used for power calculations for the *t*, chi-square and binomial tests.

Annotated outline (with Matlab scripts) for Larsen & Marx Chapter 6

6 **HYPOTHESIS TESTING**

Pierre-Simon, Marquis de Laplace (1749-1827)

6.1 **INTRODUCTION**

"The process of dichotomizing the possible conclusions of an experiment and then using the theory of probability to choose one option over the other is known as **hypothesis testing**. The two competing propositions are called the **null hypothesis** (written H_o) and the **alternative hypothesis** (written H_1 [or H_a])" A courtroom analogy is presented.

6.2 **THE DECISION RULE**

Let $y_1, y_2, ..., y_{30}$ denote the mileages recorded by each of 30 cars during a cross-country test run to evaluate a gas additive. Assume σ is known to be 2.4. Then,

$$f_Y(y;\ \mu) = \frac{1}{\sqrt{2\pi}(2.4)}\ e^{-\frac{1}{2}\left(\frac{y-\mu}{2.4}\right)^2}$$

If the existing gas mileage was 25 mpg, we are testing

H_o: μ = 25

H_1: μ > 25

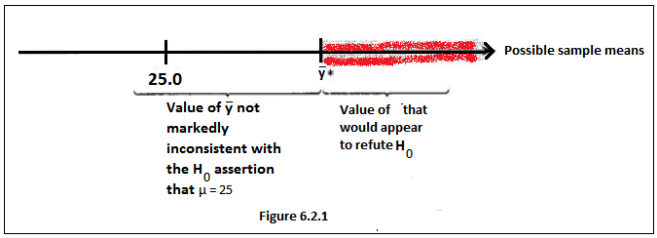

Figure 6.2.1

Is 25.25 a good choice for rejecting the null hypothesis? No. As shown in Figure 19, if the null hypothesis were true, then 28.43% of the samples of size 30 would have means ≥ 25.25.

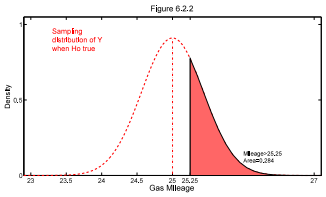

Figure 19. What percentage of means from a sample size 30 would have means ≥ 25.25 if the null hypothesis was true? 28.42%

% LMFig060202_4th.m
See **LMFig060203a_4th**

Should the cutoff be 26.5%? As shown in Figure 20, only 0.03% of the area of the curve would have means ≥ 26.5 if the null hypothesis was true (μ=25; σ=24).

% LMFig060203_4th.m
See **LMFig060203a_4th**

What would the appropriate cutoff be? We can solve for the gas mileage such that 5% or fewer of the samples would have mileages greater than that critical value. That value is 25.7207 (solved with Norminv function) and the relationship to the normal distribution is shown in Figure 21.

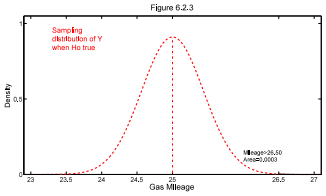

Figure 20. What percentage of means from a sample size 30 would have means ≥ 26.5 if the null hypothesis was true? .03%

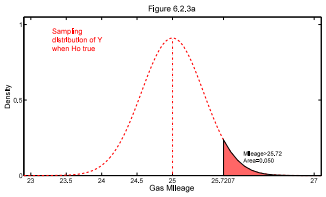

Figure 21. What percentage of means from a sample size 30 would have means ≥ 25.72 if the null hypothesis (μ=25, σ=2.4) was true? 5%

```
% LMFig060203a_4th.m
% Also Figure 6.2.2, 6.2.3, 6.2.3a, 6.4.2, 6.4.3, 6.4.6, 6.4.7
% Graphs requested by menu prompt
% Larsen & Marx (2006, p. 448). Introduction to Mathematical
Statistics
% The normal probability equation is provided on p. 264.
% This is for mean 0, and unit standard
% Written by Eugene.Gallagher@umb.edu
% written for EEOS601, written 3/10/11, revised 3/10/11
n=30;sigma=2.4;mu=25;alpha=0.05;
yhigh=1.05;
figure622=0;
figure623=0;
figure623a=0
figure642=0;
figure643=0;
figure646=0;
figure647=0;
onetailed=1;
tl='Figure 6.2.3a';
K = menu('Choose a graph','Figure 6.2.2','Figure 6.2.3',...
   'Figure 6.2.3a, 1-tailed','Figure 6.2.3a, 2-tailed','Figure 6.4.2',...
   'Figure 6.4.3','Figure 6.4.6','Figure 6.4.7');
if K==1
```

```
    tl='Figure 6.2.2';
    figure622=1;
    yup=25.25;
elseif K==2
    tl='Figure 6.2.3';
    figure623=1;
    yup=26.50;
elseif K==3
    tl='Figure 6.2.3a';
    figure623a=1;
elseif K==4
    onetailed=0;
    tl='Figure 6.2.3a';
    figure623a=1;
elseif K==5
    figure642=1;
    tl='Figure 6.4.2';
elseif K==6
    figure643=1;
    tl='Figure 6.4.3';
elseif K==7
    figure646=1;
    alpha=0.10;
    tl='Figure 6.4.6';
elseif K==8
    figure647=1;
    sigma=1.2;
    yhigh=2;
    tl='Figure 6.4.7';
end
if onetailed==1 & K>2
        yup=norminv(1-alpha,mu,sigma/sqrt(n))
        % or for those who'd like to try the optimization toolbox:
        yup2 = fsolve(@(x) (1-alpha)-normcdf((x-mu)./(sigma./sqrt(n))),25,optimset('Display','off'))
        yup-yup2  % <e-9
        fprintf('The critical value is %6.4f\n',yup)
elseif K>2
    % This is two-tailed you must solve for both tails
    yup=norminv(1-alpha/2,mu,sigma/sqrt(n));
    yup3 = fsolve(@(x) (1-alpha/2)-normcdf((x-mu)./(sigma./sqrt(n))),25,optimset('Display','off'))
    yup-yup3;
    ylow=norminv(alpha/2,mu,sigma/sqrt(n));
    % just playing around with the equation solver:
    ylow3 = fsolve(@(x) (alpha/2)-normcdf((x-mu)./(sigma./sqrt(n))),25,optimset('Display','off'))
    ylow-ylow3;  % < 10^-5
```

```
   fprintf('The lower critical value is %6.4f\n',ylow)
   fprintf('The upper critical value is %6.4f\n',yup)
end
p=1-normcdf((yup-mu)/(sigma/sqrt(n)))
muj=mu;
sigmaj=sigma/sqrt(n); % sigmaj is the standard error of the mean
miny=22.9;
if figure643==1
   maxy=28.1;
else
   maxy=27.1;
end
y=miny:.01:maxy;
fyj=normpdf(y,mu,sigmaj);
% Plot using ax1 handle, saved above,to save this graph
% on top of the previous graph.
plot(y,fyj,'linestyle','--','color','r','linewidth',3)
ylabel('Density','FontSize',20)
xlabel('Gas Mileage','FontSize',22)
axis([miny maxy 0 yhigh])
set(gca,'Ytick',[0:0.5:2],'FontSize',18)
set(gca,'Xtick',[23:.5:25 yup 27:0.5:28],'FontSize',18)
ax1=gca;% save the handle of the graph
title(tl,'FontSize',22);
hold on
fyj1=normpdf(yup,mu,sigmaj);
plot([yup yup]',[0 fyj1]','-k','linewidth',1)
fymu=normpdf(mu,mu,sigmaj);
plot([mu mu]',[0 fymu]','-.r','linewidth',1)
% Fill in the upper tail with fill
y2=yup:.1:maxy;
fyj2=normpdf(y2,mu,sigmaj);
fyjmax=normpdf(maxy,mu,sigmaj);
fill([yup y2 maxy maxy],[0 fyj2 fyjmax 0],[1 .4 .4])
if p<0.001
   t=sprintf('Mileage>%5.2f\nArea=%6.4f\n',yup,p);
else
   t=sprintf('Mileage>%5.2f\nArea=%5.3f\n',yup,p);
end
text(26,.1,t,'Color','k','FontSize',16);
if onetailed ~=1
   % Fill in the lower tail
   y4=miny:.01:ylow;
   fyj4=normpdf(y4,mu,sigmaj);
   fyjmin=normpdf(miny,mu,sigmaj);
```

```
    fill([miny miny y4 ylow],[0 fyjmin fyj4 0],[1 .4 .4])
    t=sprintf('Mileage<%5.2f\nArea=%5.3f\n',ylow,p);
    text(miny+.2,.1,t,'Color','k','FontSize',16);
end
t6=sprintf('Sampling\ndistribution of Y\nwhen Ho true');
text(23.3,.9,t6,'Color','r','FontSize',20);
if figure642==1 | figure646==1 | figure647==1
    mu2=25.750;
    fyj3=normpdf(y,mu2,sigmaj);
    % Plot using ax1 handle, saved above,to save this graph
    % on top of the previous graph.
    plot(y,fyj3,'linestyle','--','color','b','linewidth',3);
    y5=miny:.01:yup;
    fyj5=normpdf(y5,mu2,sigmaj);
    fyj5min=normpdf(miny,mu2,sigmaj);
    fyj5max=normpdf(yup, mu2,sigmaj);
    fill([miny miny y5 yup],[0 fyj5min fyj5 0],[.8 .8 1])
    t=sprintf('Sampling\ndistribution of Y\nwhen mu=%5.2f',mu2);
    text(26.15,.9,t,'Color','b','FontSize',20);
    fyj6=normpdf(yup,mu2,sigmaj);
    plot([yup yup]',[0 fyj6]','-b','linewidth',1)
    h1 = findobj(gca,'Type','patch');
    set(h1,'facealpha',0.5);
    Beta=normcdf(yup,mu2,sigmaj);
    t5=sprintf('Beta=%6.4f\n',Beta);
    text(25.05,.05,t5,'Color','b','FontSize',16);
    fymu6=normpdf(mu2,mu2,sigmaj);
    plot([mu2 mu2]',[0 fymu6]','-.b','linewidth',1)
elseif figure643==1
    mu2=26.8;
    fyj3=normpdf(y,mu2,sigmaj);
    % Plot using ax1 handle, saved above,to save this graph
    % on top of the previous graph.
    plot(y,fyj3,'linestyle','--','color','g','linewidth',3);
    y5=miny:.01:yup;
    fyj5=normpdf(y5,mu2,sigmaj);
    fyj5min=normpdf(miny,mu2,sigmaj);
    fyj5max=normpdf(yup, mu2,sigmaj);
    fill([miny miny y5 yup],[0 fyj5min fyj5 0],[.8 .8 1])
    t=sprintf('Sampling\ndistribution of Y\nwhen mu=%5.2f',mu2);
    text(27,.9,t,'Color','g','FontSize',20);
    fyj6=normpdf(yup,mu2,sigmaj);
    plot([yup yup]',[0 fyj6]','-g','linewidth',1)
    h1 = findobj(gca,'Type','patch');
    set(h1,'facealpha',0.5);
```

```
   Beta=normcdf(yup,mu2,sigmaj);
   t5=sprintf('Beta=%6.4f\n',Beta);
   text(24.6,.05,t5,'Color','g','FontSize',16);
   fymu6=normpdf(mu2,mu2,sigmaj);
   plot([mu2 mu2]',[0 fymu6]','-.g','linewidth',1)
end
figure(gcf)
hold off
```

Table 6.2.1

```
% LMTable060201_4th.m
% from the normal pdf
% From Larsen & Marx  (2006) Introduction to Mathematical Statistics,
% Fourth Edition. page 432
% Dept. Environmental, Earth & Ocean Sciences
% Written by Eugene.Gallagher@umb.edu written & revised 3/4/11
% http://alpha.es.umb.edu/faculty/edg/files/edgwebp.html
% Use Matlab to find the 95% critical value (1-tailed)? it is about 25.178
y1 = fsolve(@(x) 0.95-normcdf((x-25)./(2.4./sqrt(30))),25,optimset('Display','off'))
% Generate n random samples of size 30 with mean 25 and sigma=2.4
MU=25;SIGMA=2.4;
fprintf('mu =%5.3f and sigma = %5.3f\n',MU,SIGMA);
n=75;                  % Larsen & Marx used 75
samsize=30;            % Larsen & Marx used 30
Ybar = mean(normrnd(MU,SIGMA,samsize,n));
fprintf('\nTable 6.2.1\n\n')
for i=1:n
   if Ybar(i)>=y1
      fprintf('%6.3f \tyes\n',Ybar(i))
   else
      fprintf('%6.3f \tno\n',Ybar(i))
   end
end
fprintf('There were %1.0f of %2.0f samples that exceeded %6.3f.\n',sum(Ybar>=y1),n,y1)
```

Comment (p. 433). If H_o: $\mu=\mu_o$ is rejected using a 0.05 **decision rule**, we say that the difference between $\bar{y}$ and μ_o is **statistically significant** at the 5% **α-level**.

6.2.1 Expressing decision rules in terms of Z ratios

Z ratio isn't defined, but it is any ratio that is expected to be distributed as the standard normal distribution (μ = σ = 1). Z ratios are usually formed by as the ratio of {a statistic - expected value} divided by the standard error of the statistic.

Rejecting H_o: μ=25.0 when

$$\bar{y} \geq \bar{y}^* = 25 + norminv(.95) \cdot \frac{2.4}{\sqrt{30}} = 25.0 + 1.6449 \cdot \frac{2.4}{\sqrt{30}} = 25.7207$$

rejecting H_o when $\dfrac{\bar{y} - 25.0}{2.4/\sqrt{30}} \geq 1.64$.

Definition 6.2.1 Any function of the observed data whose numerical value dictates whether H_o is accepted or rejected is called a **test statistic**. The set of values for the test statistic that result in the null hypothesis being rejcted is called the **critical region** and is denoted *C*. The particular point in C that separates the rejection region from the acceptance region is called the **critical value**.

Definition 6.2.2 The probability that the test statistic lies in the critical region when H_o is true is called the **level of significance** and is denoted α. [significance level is also called **α-level**]

6.2.2 One Sided Versus Two-Sided Alternatives

"If there is reason to believe before any data are collected that the parameters being tested is necessarily restricted to one particular "side" of H_o, then H_1 is defined tgo reflect that limitation and we say that the alternative **hypothesis is one-sided**... If no such a priori information is available, the alternative hypothesis needs to accommodate the possibility that the true parameter value might lie one either side of μ_o. Any such alternative is said to be **two-sided**. For testing H_o: μ=μ_o, the two sided alternative is written H_1: $\mu \neq \mu_o$." (p 434)

6.2.3 Testing H_o: μ=μ_o (σ known)

Figure 6.2.4

```
% LMFig060204_4th.m
% Larsen & Marx (2001, p. 367).
% Normal probability pdf
% Written by Eugene.Gallagher@umb.edu
% written for EEOS601
z05=norminv(0.95)
z=-3.05:.001:3.05;
fzz=normpdf(z)
plot(z,fzz,'linestyle','--','color','r','linewidth',3)
ylabel('f_z(z)','FontSize',20)
xlabel('z','FontSize',22)
axis([-3.05 3.05 0 0.405])
set(gca,'Ytick',[0 0.2 0.4],'FontSize',18)
```

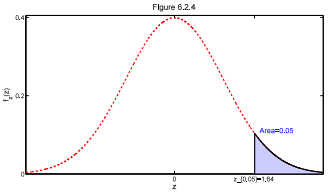

Figure 22. The z score for a 1-tailed critical value for α = 0.05 is 1.64 (=norminv(1-0.05)).

```
set(gca,'Xtick',[0 1.64],'XtickLabels',{'0','z_{0.05}=1.64'},'FontSize',18)
ax=axis;
ax1=gca; % save the handle of the graph
title('Figure 6.2.4','FontSize',22)
hold on
fz05=normpdf(z05);
plot([z05 z05]',[0 fz05]','-k','linewidth',1)
% Fill in the upper tail with fill
y2=z05:.001:ax(2);
fy2=normpdf(y2);
fymax=normpdf(ax(2));
fill([z05 y2 ax(2) ax(2)],[0 fy2 fymax 0],[.8 .8 1])
t=sprintf('Area=%4.2f\n',p);
text(z05+.1,.1,t,'Color','b','FontSize',20);
figure(gcf)
hold off
```

Theorem 6.2.1 (1-sample z test) Let y_1, y_2, ..., y_n be a random sample of size n from a normal distribution where σ is known. Let $z = \dfrac{\bar{y} - \mu_o}{\sigma/\sqrt{n}}$.

a. To test H_o: $\mu = \mu_o$ versus H_1: $\mu > \mu_o$ at the α level of significance, reject H_o if $z \geq z_\alpha$.
b. To test H_o: $\mu = \mu_o$ versus H_1: $\mu < \mu_o$ at the α level of significance, reject H_o if $z \leq -z_\alpha$.
c. To test H_o: $\mu = \mu_o$ versus H_1: $\mu \neq \mu_o$ at the α level of significance, reject H_o if z is either (1) $\leq -z_{\alpha/2}$ or (2) $\geq z_{\alpha/2}$.

Example 6.2.1

```
% LMex060201_4th.m
% Larsen & Marx (2001, p. 367).
% Introduction to Mathematical Statistics, 4th edition
% Normal probability pdf
% Written by Eugene.Gallagher@umb.edu
% written for EEOS601
z025=norminv(0.025);
z975=norminv(0.975);
z=-3.05:.001:3.05;
fzz=normpdf(z)
plot(z,fzz,'linestyle','--','color','r','linewidth',3)
ylabel('f_z(z)','FontSize',20)
xlabel('z','FontSize',22)
axis([-3.05 3.05 0 0.405])
set(gca,'Ytick',[-1.96 0 0.6 1.96],'FontSize',18)
set(gca,'Xtick',[-1.96 0 0.60 1.96],...
   'XtickLabel',{'-1.96', '0','z=0.60','1.96'},'FontSize',18)
ax=axis;
```

```
ax1=gca; % save the handle of the graph
title('Figure 6.2.5','FontSize',22)
hold on
fz=normpdf([z025 z975]);
plot([z025 z025;z975 z975]',[0 fz(1);0 fz(2)]','-k','linewidth',1)
% Fill in the upper tail with fill
y2=z975:.001:ax(2);
fy2=normpdf(y2);
fymax=normpdf(ax(2));
fill([z975 y2 ax(2) ax(2)],[0 fy2 fymax 0],[.8 .8 1])
% Fill in the lower tail with fill
y3=ax(1):.001:z025;
fy3=normpdf(y3);
fymin=normpdf(ax(1));
fill([ax(1) ax(1) y3 z025],[0 fymin fy3 0],[.8 .8 1])
t=sprintf('Area=0.025');
text(ax(1)+.3,.2,t,'Color','b','FontSize',20);
text(1.6,.2,t,'Color','b','FontSize',20);
figure(gcf)
hold off
```

6.2 The P-value

Definition 6.2.3 The ***p-value*** associated with an observed test statistic is the probability of getting a value for that test statistic as extreme or more extreme than what was actually observed (relative to H_1) *given that H_o is true*.

Example 6.2.2
[Solved as part of LMex060201_4th.m

Questions p 438-439

6.3 Testing binomial data - $H_o: p=p_o$ (p 440)

6.3.1 A Large Sample Test for the Binomial Parameter p

Theorem 6.3.1 Let $k_1, k_2, \ldots, k_n$ be a random sample of n Bernoulli random variables for which $0 < n\,p_o\,3\sqrt{(np_o\,(1-p_o))} < np_o + 3\sqrt{(np_o\,(1-p_o)} < n$. Let $k = k_1 + k_2 + \ldots + k_n$ denote the total number of "successes" in the n trials. Define $z = (k-np_o)/(\sqrt{(np_o(1-p_o))}$,

to test $H_o: p = p_o$ versus $H_1: p > p_o$ at the α level of significance, reject H_o if $z \geq z_\alpha$.
to test $H_o: p = p_o$ versus $H_1: p < p_o$ at the α level of significance, reject H_o if $z \geq -z_\alpha$.
to test $H_o: p = p_o$ versus $H_1: p \neq p_o$ at the α level of significance, reject H_o if z is either (1) $\leq -z_{\alpha/2}$ or (2) $\geq z_{\alpha/2}$.

Case Study 6.3.1 Point Spread
Point spreads are designed so that the both teams should have an equal probability of winning once the point spread has been added to the team's points. In a study of 124 NFL games, the favored team beat the spread in 67 games (54%) of the time. Is this more than could be expected by chance? Apply the one sample binomial test. (See text above for fuller description)

Passed the test for the normal approximation
Using the normal approximation for the 1-sample binomial, z=0.8980 and 2-tailed p=0.3692

Using the binomial pdf for the 1-sample binomial, 2-tailed p=0.4191
Using the binomial pdf for the 1-sample binomial, 1-tailed p=0.2095

```
% LMcs060301_4th.m
% Larsen & Marx Case Study 6.3.1 (2006, p. 441)
% Point Spread example
% Point spreads are set so 50% of teams beat the spread
% During one period 67 of 124 teams beat the spread
% What is the probability that 67/124 could have been
% observed by chance if the true p of beating the spread
% is 0.5?
X=67;
n=124;
p=0.5;
% Check Demoivre-Laplace rule, Larsen & Marx (2006, Equation 6.3.1)
OUT=demoivre(124,p);
if OUT==1
   fprintf('The DeMoivre-Laplace rule is met, Normal approximation probably ok\n');
else
   fprintf('The DeMoivre-Laplace rule not met, Normal approximation not ok\n');
end
usebinomial=0;  % Will perform the approximate test
[p2tailed,exact,zval]=onesamplebinom(n,X,p,usebinomial);
fprintf(...
   'Using the normal approximation for the 1-sample binomial, z=%6.4f and 2-tailed
p=%6.4f\n',...
   zval,p2tailed);
usebinomial=1;
[p2tailed,exact]=onesamplebinom(n,X,p,usebinomial);
fprintf(...
   'Using the binomial pdf for the 1-sample binomial, 2-tailed p=%6.4f\n',...
   p2tailed);
UpperTailp=binomutp(n,X,p);
fprintf(...
   'Using the binomial pdf for the 1-sample binomial, 1-tailed p=%6.4f\n',...
   UpperTailp);
```

```
function [p2tailed,exact,zval]=onesamplebinom(n,k,po,usebinomial);
% format[pvalue,exact,zval]=onesamplebinom(n,k,po,usebinomial);
% Input n=number of cases
%       k=number of successes
%       po=expected proportion
%       usebinomial=1 for exact test [optional]
% Output: p2tailed  2-tailed p value
%         exact     1 if Exact binomial test used
%         zval = z statistic
% calls demoivre.m
```

```
% Reference: Larsen & Marx (2006, p. 440) Theorem 6.3.1
% Uses bernoull.m
if nargin<4 | usebinomial==0
   usebernoull=logical(0);
else
   usebernoull=logical(1);
   exact=1;
end
if ~usebernoull
   % Equation 6.3.1 in L&M 4th edition p. 631
  if 0 < n*po-3*sqrt(n*po*(1-po)) & n*po-3*sqrt(n*po*(1-po)) < ...
   n*po +3*sqrt(n*po*(1-po)) & n*po +3*sqrt(n*po*(1-po)) < n;
      fprintf('Passed the test for the normal approximation\n')
      exact=logical(0);
   else
      fprintf('Failed test for normal approximation, Must use the binomial test\n')
      exact=logical(1);
   end
end
Expected=n*po;
Varpo=po*(1-po);
zval=(k-Expected)/sqrt(n*Varpo);
if exact
   Expectedk=n*po;
   dev=abs(Expectedk-k);
   if Expectedk>=k
      i=0:k;
      lowertailp=sum(binopdf(i,n,po));
      i=ceil(Expectedk+dev):n;
      uppertailp=sum(binopdf(i,n,po));
      p2tailed=lowertailp+uppertailp;
   else
      i=k:n;
      uppertailp=sum(binopdf(i,n,po));
      i=0:floor(Expectedk-dev);
      lowertailp=sum(binopdf(i,n,po));
      p2tailed=lowertailp+uppertailp;
   end
else
   if zval<0
       p2tailed=2*(normcdf(zval));
   elseif zval>=0
      p2tailed=2*(1-normcdf(zval));
   end
end
```

```
function UpperTailP=binomutp(n,k,p);
% format UpperTailp=binomutp(n,k,p);
% Input n=number of cases
%       k=number of successes
%       p=expected proportion
% Output: UpperTailP Upper-Tail probabilities.
% Reference: Hollander & Wolfe p. 567.
% written by Eugene.Gallagher@umb.edu
if k>n
   error('k must be <= n');
end
i=k:n;
UpperTailP=sum(binopdf(i,n,p));
```

Case Study 6.3.2 Deaths after birthdays

6.3.2 A small sample test for the binomial parameter p

Example 6.3.1

```
% LMex060301_4th.m
% page 444-445 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Written by Eugene.Gallagher@umb.edu, 2003, revised 3/8/11
p=0.85;
n=19;
Expected=p*n
k=0:19;
pdf=binopdf(k,n,p);
i=find(k<Expected);
alpha=0.1;
fprintf(...
   'Goal is to find 2-tailed critical values for alpha=%4.2f\n',...
   alpha);
cpdf=cumsum(pdf(i));
disp(['k           ';'pdf         ';'cumulative pdf'])
disp([k(i);pdf(i);cpdf]);
j=max(find(cpdf<=alpha/2));
if cpdf(j)~=alpha/2
   j=j+1;
end
fprintf('The lower critical value is %d with cumulative pdf of %6.4f\n',...
   k(i(j)),cpdf(j));
fprintf('Find the upper critical value\n')
i=find(k>Expected);
ucpdf=fliplr(cumsum(fliplr(pdf(i))));
disp([k(i);pdf(i);ucpdf]);
m=min(find(ucpdf<=alpha/2));
% By inspecting the cumulative pdf, the goal of attaining a nearly
```

```
% symmetric set of critical values around the expected values is
% met by setting the upper critical value at 19
fprintf('The upper critical value is %d with cumulative pdf of %6.4f\n',...
   k(i(m)),ucpdf(m));
fprintf('With these critical values & Ho: p=%4.2f, P(Type I error)=alpha=%6.4f\n',...
   p,ucpdf(m)+cpdf(j))
```

Questions p 445-446

6.4 TYPE I AND TYPE II ERROR

Type I and Type II error There are two kinds of errors that can be committed in the process of hypothesis testing. They are shown in the following table. **Type I error** is the probability of rejecting the null hypothesis H_o when H_o is true; **Type II error** is the probability of failing to rject the null hypothesis H_o when H_o is false.

Table 2. Hypothesis testing decision tree, Type I and Type II errors. **Larsen & Marx (2006, p 447)**.

		True State of Nature	
		H_o is true	H_1 is true
Our Decision	Fail to reject H_o	Correct Decision	Type II error
	Reject H_o	Type I error	Correct Decision

6.4.1 Computing the Probability of Committing a Type I Error

Recall the fuel additive example in Section 6.2. With μ=25 & σ=2.4, the **decision rule** stated that H_o should be rejected at an α-level of 0.05 if $\bar{y}$ equaled or exceeded 25.7207 (25.781 due to rounding in the text. The probability of committing Type I error is set by the **decision rule** at 0.05:

$$P(\text{Type I error}) = P(\text{reject } H_o \mid H_o \text{ is true})$$
$$= P(\bar{y} \geq 25.7207 \mid \mu = 25 \ \& \ \sigma = 2.4)$$
$$= P\left(\frac{\bar{Y} - 25\ mpg}{2.4\ mpg/\sqrt{30}} \geq \frac{25.7207\ mpg - 25\ mpg}{2.4\ mpg/\sqrt{30}} \right)$$
$$P(Z \geq 1.64) = 0.05$$

The probability of committing a Type I error is a test's **level of significance** (recall **Definition 6.2.2**). "The concept is a crucial one: The level of significance is a single number summary of the "rules" by which the decision process is being conducted. In essence α reflects the amount of evidence the experimenter is demanding to see before abandoning the null hypothesis." p. 448

6.4.2 Computing the Probability of Committing a Type II Error

6.4.2.1 Type II error probabilities are not explicitly set by the experimenter

6.4.2.2 Each hypothesis has an infinite number of Type II probabilities, one for each parameter admissable under H_1.

For example, what is the probability of committing a Type II error in the gasoline experiment if μ_o were 25 mpg, but the true μ (with the additive) were 25.750. By definition,

$$P(\text{Type II error} \mid \mu = 25.750) = P(\text{fail to reject } H_o \mid \mu = 25.750 \ \& \ \sigma = 2.4)$$

$$= P(\bar{Y} < 25.781 \mid \mu = 25.750 \ \& \ \sigma = 2.4)$$

$$= P\left(\frac{\bar{Y} - 25.75}{2.4/\sqrt{30}} < \frac{25.718 - 25.75}{2.4/\sqrt{30}} \right)$$

$$P(Z < -0.07) = 0.4721$$

So, even if the fuel additive increased gas mileage to 25.75 mpg, our **decision rule** would be tricked 47.21% of the time, telling us not to reject H_o. The symbol for the probability of committing type II error is β.

Figure 23 shows the sampling distribution of $\bar{y}$ when $\mu = 25$ & $\sigma = 2.4$ and when $\mu = 25.75$ & $\sigma = 2.4$ (H_1 is true).

% LMFig060402_4th.m
See LMFig060203a_4th.m

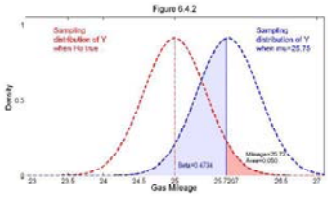

Figure 23. If μ=25.75, a decision rule based on H_o: μ=25.0, σ=2.4 & n=30 (i.e., critical value =25.2707= norminv(.95,25,2.4/sqrt(30))) would have a P (Type II error) = β = 0.4734 = normcdf(norminv(.95,25,2.4/sqrt(30)),25.75,2.4/sqrt(30))

Figure 24 shows the probability of Type II error if μ = 26.8 mpg. See LMFig060203a_4th.m for program.

6.4.3 Power Curves

"If β is the probability that we fail to reject H_o when H_1 is true, then 1 - β is the probability of the complement, that we reject H_o when H_1 is true. We call 1 - β the **power** of the test; it represents the ability of the **decision rule** to "recognize" (correctly) that H_o is false." A **power curve** is a graph of 1 - β versus the set of all possible parameter values.

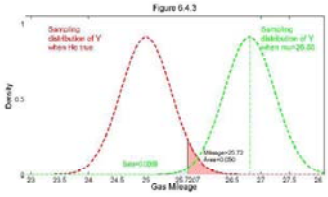

Figure 24. If μ=26.8, a decision rule based on H_o: μ=25.0, σ=2.4 & n=30 (i.e., critical value =25.2707= norminv(.95,25,2.4/sqrt(30))) would have a P (Type II error) = β = 0.0069 = normcdf(norminv(.95,25,2.4/sqrt(30)),26.8,2.4/sqrt(30))

Figure 25 shows the power curve for testing H_o: μ = 25.0 & σ = 2.4 vs. H_1: μ > 25.0 & σ = 2.4.

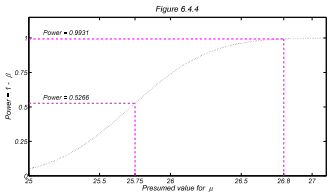

Figure 25. Power curve for H_o = 25, σ = 2.4 & n=30. Also shown are the estimated power of the test against alternative hypotheses μ = 25.75 and μ = 26.8

```
% LMFig060404_4th.m
% Using as a model %LMcs030201_4th.m, page 450 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics,
4th edition
% Written 3/10/11 (11:22 -)
% Eugene.Gallagher@umb.edu for EEOS601 UMASS/Boston
CritVal=norminv(.95,25,2.4/sqrt(30));
xmin=25;xmax=27.1;
X=xmin:0.005:xmax;
sigmaj=2.4;n=30;
Power=1-normcdf(CritVal,X,sigmaj/sqrt(n));
x=[25.75 26.8]; Pow=1-normcdf(CritVal,x,sigmaj/sqrt(n));
plot(X,Power,':k','LineWidth',3)
axis([xmin xmax 0, 1.1])
hold on
set(gca,'Xtick',[25 25.5 x(1) 26 26.5 x(2) 27],'Ytick',[0:.25:1],'FontSize',18)
% This plots horizontal and vertical lines on the graph
plot([x(1) x(1);xmin x(1);x(2) x(2);xmin x(2)]',[0 Pow(1);...
   Pow(1) Pow(1);0 Pow(2);Pow(2) Pow(2)]','--m','LineWidth',3);
ylabel ('Power = 1 - Beta','FontSize',20);
xlabel('Presumed value for mu','FontSize',20)
title('Figure 6.4.4','FontSize',22)
t1=sprintf('Power = %6.4f',Pow(1));
t2=sprintf('Power = %6.4f',Pow(2));
text(xmin+0.2,0.55,t1,'FontSize',18);
text(xmin+0.2,1.02,t2,'FontSize',18);
figure(gcf)
pause
hold off
```

Figure 26 shows the power curves for the gas mileage experiment comparing a 60-car and 30-car experiment. The power curves are shown for a 2-tailed test.

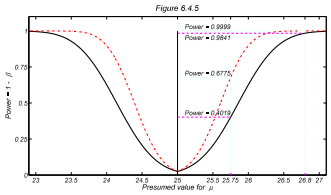

Figure 26. Two-tailed power curves for H_o = 25, σ = 2.4 & n=30 (black) and n=60 (red dashed lines). Also shown are the estimated power of each test against alternative hypotheses μ = 25.75 and μ = 26.8. With n=30, the power of the test against H_1: μ=25.75 is 0.4019, but with n=60, the power increases to 0.6775. The power also increases versus H_1: μ=26.8 from 0.9841 to 0.9989. The previous figure was the power curve for the 1-tailed test, which is more powerful than the 2-tailed test for equal sample size.

6.4.4 Factors That Influence the Power of a Test

6.4.4.1 The effect of α on 1 - β

Figure 27 shows what happens to 1-β (when μ = 25.75) if σ, n, and μ are held constant but α is increased to 0.10 instead of 0.05.

Figure 6.4.6

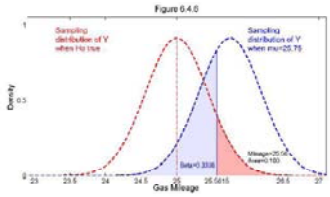

Figure 27. Increasing α from 0.05 to 0.10 decreases β from 0.4734 (see Figure 6.4.2) to 0.336, thus increasing power from 53% to 67%.

6.4.5 The effects of σ and *n* on 1- β

Figure 28 reveals the strong effects of reducing σ on the power of a test.

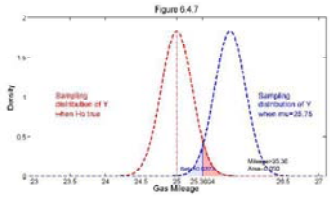

Figure 28. Decreasing σ from 2.4 to 1.2 decreases β from 0.4734 (see Figure 6.4.2) to 0.0377, thus increasing power from 53% to 96%.

Example 6.4.1

```
% LMEx060401_4th.m
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th ed
% Written by Eugene.Gallagher@umb.edu, 3/10/11 revised 3/10/11
muo=100;sigma=14;Power=.6;muh1=103;alpha=0.05
N = sampsizepwr('z',[muo sigma],muh1,Power,[],'alpha',alpha,'tail','right');
fprintf('%2.0f samples must be taken to achieve power=%2.0f%%\n',N,Power*100);
% Optional, just checking why L&M said 78 was adequate
% What is th exact value of n needed; use an anonymous function to solve n
initialguess=75
x=fsolve(@(n) 100+norminv(0.95)*14/sqrt(n)-(103-0.25*14/sqrt(n)),...
   initialguess,optimset('Display','off'));
fprintf('The exact sample size required is %6.4f\n',x);
% Larsen & Marx report 78 so check the power with 78.
n=78;
Power = sampsizepwr('z',[muo sigma],muh1,[],n,'alpha',alpha,'tail','right');
fprintf('One-tailed test: the power with n=%2.0f is %5.2f%%.\n',n,Power*100)
% Another way to calculate power, should be identical
CriticalValue=norminv(1-alpha,muo,sigma/sqrt(n));
Power2=1-normcdf(CriticalValue,muh1,sigma/sqrt(n));
fprintf('One-tailed test: the power with n=%2.0f is %5.2f%%.\n',n,Power2*100)
% note that if the test were two-tailed, here is how these functions would
% be called:
Power = sampsizepwr('z',[muo sigma],muh1,[],n,'alpha',alpha,'tail','both');
fprintf('Two-tailed test: the power with n=%2.0f is %5.2f%%.\n',n,Power*100);
CriticalValue2=norminv(1-alpha/2,muo,sigma/sqrt(n));
Power2=1-normcdf(CriticalValue2,muh1,sigma/sqrt(n));
fprintf('Two-tailed test: the power with n=%2.0f is %5.2f%%.\n',n,Power2*100)
```

6.4.6 Decision Rules for Nonnormal data

Example 6.4.2 Skipped, life's too short

Example 6.4.3 Not programmed but doable
Four measurements — k1, k2, k3 & k4 — are taken on a Poisson random variable, X for the purpose of testing H_o: λ=0.8 vs. λ > 0.8 What is the decision rule if the level of significance is to be 0.10 and what will the power of the test be when λ = 1.2?

```
% LMex060403_4th.m
% Page 457-458 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics, 4th edition
% Application of poisspdf
% Four measurements — k1, k2, k3 & k4 — are taken on a Poisson random
% variable, X for the purpose of testing Ho: ?=0.8 vs. ? > 0.8
% A) What is the decision rule if the level of significance is to be 0.10?
% B) What will the power of the test be when ? = 1.2?
% For A, the expected value of 4 observations is 4*lambda (Example 31210),
% So analyze the pdf for lambda=4*0.8
% Written in 2003 by Eugene.Gallagher@umb.edu, revised 3/11/2011
lambda=4*0.8;alpha=0.1;
% Right tailed test
X=0:20;
% for a right-tailed test:
testcdf=1-poisscdf(X,lambda);
% for a left tailed test use: testcdf=poisscdf(X,lambda);
i=min(find(testcdf<alpha));
fprintf('The critical value for alpha=%3.1f is %2.0f\n',alpha,X(i));
fprintf('With crtical value = %2.0f, the alpha level is %6.4f\n',...
   X(i),testcdf(i))
% B) What is the power of the test if lambda= 1.2?
lambdah1=1.2;
% Right tailed test, appropriate for this problem:
PTypeII=poisscdf(X(i-1),lambdah1*4);
% Left tailed test: %PTypeII=1-poisscdf(X(i-1),lambdah1*4)
Power=1-PTypeII;
fprintf('With critical value=%2.0f & lambda=%3.1f, P(Type II error)=%6.4f, and the power of
the test is %5.2f%%\n',...
   X(i),lambdah1,PTypeII,Power*100)
% Power is 0.349
```

Example 6.4.4 Not programmed but doable

Questions

6.5 A NOTION OF OPTIMALITY: THE GENERALIZED LIKELIHOOD RATIO *[Summer 2011 students can skip this section]*

Generalized likelihood ratio
Definition 6.5.1
Definition 6.5.2
Questions p. 465-

6.6 TAKING A SECOND LOOK AT STATISTICS (STATISTICAL SIGNIFICANCE VERSUS "PRACTICAL" SIGNIFICANCE)

Figure 29 shows the 1-tailed power curve for the gasoline additive c

Figure 30 shows the 1-tailed power curves for the gasoline additive data for sample sizes of 30, 60 and 900.

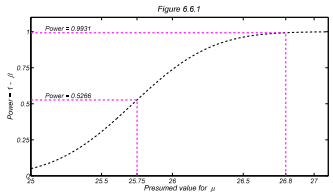

Figure 29. One-tailed power curve for H_o = 25, σ = 2.4 & n=30. Also shown is the estimated power of the test against alternative hypotheses μ = 25.75 and μ = 26.8. With n=30, the power of the 1-tailed test against H_1: μ=25.75 is 0.5266. The two-tailed test shown in a previous figure had a power of only 0.4019. The power of the 1 tailed test against H_1: μ=26.8 is 99.31%, which is an increase from 98.41%. Tests against 1-tailed alternatives are more powerful than two-tailed tests.

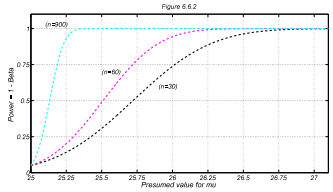

Figure 30. One-tailed power curves for H_o = 25, σ = 2.4 & n=30, 60 and 900.

```
%LMFig060601_4th.m
% Also plots Figure 6.6.2
% Based on LMFig060405_4th.m
% LMFig060405_4th.m
% Using as a model %LMcs030201_4th.m, page 450 in
% Larsen & Marx (2006) Introduction to Mathematical Statistics,
4th edition
% Written 3/10/11 (11:22 -11:57 am, )
% Eugene.Gallagher@umb.edu for EEOS601 UMASS/Boston
xmin=25;xmax=27.1;mu=25;
X=mu:0.005:xmax;
sigmaj=2.4;n=30;alpha=0.05;
K = menu('Choose a graph','Figure 6.6.1','Figure 6.6.2')
CritVal=norminv(1-alpha,mu,sigmaj/sqrt(n));
Power =1-normcdf(CritVal,X,sigmaj/sqrt(n));
x=[25.75 26.8];
Pow=1-normcdf(CritVal,x,sigmaj/sqrt(n));
plot(X,Power,'--k','LineWidth',3)
axis([xmin xmax 0, 1.1])
ylabel ('Power = 1 - Beta','FontSize',20);
xlabel('Presumed value for mu','FontSize',20)
hold on
if K==1
   % This plots horizontal and vertical lines on the graph
   plot([mu mu]',[0 1]','k','LineWidth',2)
   plot([x(1) x(1);mu x(1);x(2) x(2);mu x(2)]',[0 Pow(1);...
   Pow(1) Pow(1);0 Pow(2);Pow(2) Pow(2)]','--m','LineWidth',3);
   title('Figure 6.6.1','FontSize',22)
   t1=sprintf('Power = %6.4f',Pow(1));
   t2=sprintf('Power = %6.4f',Pow(2));
   text(mu+0.1,Pow(1)+.03,t1,'FontSize',18);
   text(mu+0.1,Pow(2)+.03,t2,'FontSize',18);
   set(gca,'Xtick',[23:0.5:25.5 x(1) 26 26.5 x(2) 27],'Ytick',[0:.25:1],'FontSize',18)
else
   n2=60;n3=900;
   CritVal2=norminv(1-alpha,mu,sigmaj/sqrt(n2));
   CritVal3=norminv(1-alpha,mu,sigmaj/sqrt(n3));
   Power2 =1-normcdf(CritVal2,X,sigmaj/sqrt(n2));
   Power3 =1-normcdf(CritVal3,X,sigmaj/sqrt(n3));
   Pow2=1-normcdf(CritVal2,x,sigmaj/sqrt(n2));
   Pow3=1-normcdf(CritVal3,x,sigmaj/sqrt(n3));
```

```
    plot(X,Power2,'--m','LineWidth',3)
    plot(X,Power3,'--c','LineWidth',3)
    title('Figure 6.6.2')
    t1=sprintf('(n=%2.0f)',n);
    t2=sprintf('(n=%2.0f)',n2);
    t3=sprintf('(n=%3.0f)',n3)
    text(25.9,0.6,t1,'FontSize',18);
    text(25.5,0.7,t2,'FontSize',18);
    text(25.1,1.03,t3,'FontSize',18);
    set(gca,'Xtick',[23:0.25:27],'Ytick',[0:.25:1],'FontSize',18)
    grid
end
figure(gcf)
pause
hold off
```

References

Berger, J. O. and D. A. Berry. 1988. Statistical analysis and the illusion of objectivity. American Scientist *76*: 159-165. {**6**}

Brown, L. D., T. T Cai, and A. DasGupta. 2001. Interval estimation for a binomial proportion. Statistical Science *16*: 101-133.{**20**}

Dennis, B. 1996. Should ecologists become Bayesians? Ecological Applications *6*: 1095-1103.{**5**, **6**}

Huey, R. B, G. W. Gilchrist, M. L. Carlson, D. Berrigan and L. Serra. 2000. Rapid evolution of a geographic cline in size in an introduced fly. Science *287*: 308-309.{**8**, **9**}

Ioannidis, J. P. A. 2005. Why most published research findings are false. PloS Medicine *2*: 696-701. {**11**}

Larsen, R. J. and M. L. Marx. 2006. An introduction to mathematical statistics and its applications, 4th edition. Prentice Hall, Upper Saddle River, NJ. 920 pp. {**7**, **9**, **10**, **11**, **12**, **20**, **21**, **33**}

Mayo, D. G. 1996. Error and the growth of experimental knowledge. University of Chicago Press, Chicago. 493 pp. {**5**, **6**, **8**}

Ramsey, F. L. and D. W. Schafer. 2002. The statistical sleuth: a course in methods of data analysis, 2nd Edition. Duxbury Press, Belmont CA, 742 pp & data CD. {**9**, **10**}

Salsburg, D. 2001. The lady tasting tea: how statistics revolutionized science in the twentieth century. W. H. Freeman & Co., New York. 340 pp. *[This is a wonderful little book, with superb New Yorker style articles on Fisher, Neyman, Kolmogorov, Wilcoxon, Deming and other giants of 20th century statistics]* {**5**, **6**}

Sterne, J. A. C. and G. D. Smith. 2001. Sifting the evidence — what's wrong with significance tests? British Medical Journal *322*: 226-231. *[Available online with free registration at: **http://bmj.bmjjournals.com/cgi/reprint/322/7280/226**]{**4**, **6**, **9**, **10**, **11**}*

Suppes, P. 1969. Models of data. Pp 24-35 in Studies in the methodology and foundations of science. Dordrecht, The Netherlands. [Cited by **Mayo (1996)** but unavailable in Google Scholar]{**8**}

Index