Extra Material, not in the text

Theorem 10.3b Let $T_1, T_2, \ldots$ be a sequence of random variables. If $\lim_{n\to\infty} E(T_n) = \theta$ and $\lim_{n\to\infty} Var(T_n) = 0$ then $T_n \xrightarrow{P} \theta$.

This is proved using Markov's inequality, with $g(x) = (x-\theta)^2$ and $a = c^2$.

Continuous mapping theorem Let $\mathbf{T}_n = (T_n^{(1)}, \ldots T_n^{(k)})'$, $\mathbf{t}_0 = (t_0^{(1)}, \ldots t_0^{(k)})'$, and $T_n^{(j)} \xrightarrow{P} t_0^{(j)}$ for $j = 1, \ldots, k$. If the function g is continuous at $\mathbf{t}_0$, then $g(\mathbf{T}_n) \xrightarrow{P} g(\mathbf{t}_0)$.

Method of Moments The description of this method that is presented in the text (and initially, what I said in class) applies just to random samples. But the following version is a bit more general. It is useful for data that are not identically distributed, like regression data. In this formulation, the data need not even be independent.

Let the joint distribution of the data $X_1, \ldots, X_n$ depend on the parameters $\theta_1, \ldots \theta_r$. Following standard notation, we define the kth sample moment as

$$m'_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

Clearly, $E(m'_k)$ is a function of $\theta_1, \ldots \theta_r$. So pick r values of k, and write the system of equations

$$\begin{aligned} E(m'_{k_1}) &= g_1(\theta_1, \ldots \theta_r) \\ &\vdots \\ E(m'_{k_r}) &= g_r(\theta_1, \ldots \theta_r) \end{aligned}$$

Once this has been done, follow these steps.

1. Remove the expected values on the left hand side of the equations.
2. Add hats to the θ values on the right hand side.
3. Solve for the $\widehat{\Theta}$ values. These are the Method of Moments (MOM) estimators.

Here are some comments.

- If the data are a random sample (independent and identically distributed) then $E(m'_k) = \mu'_k$, and we have the usual Method of Moments estimators.
- Of course it does not matter if you solve the equations first and then put the hats on, or if you do it the other way around.
- You can hope that the values of k are just $1, \ldots, r$, but there is no guarantee. In particular, if the distribution of the X_i values is symmetric about zero, then all the odd-numbered moments are zero, and only even values of k will be useful.
- If the data are not a random sample (think regression), don't depend on formulas like $E(\overline{X}) = \mu$; they were derived for the *i.i.d.* case, and might not be correct. Do the calculation.

An i.i.d Example Let $X_1, \ldots, X_n$ be a random sample from a Binomial distribution with parameters k and θ, *both* unknown. We write the system of equations

$$\begin{aligned} E(m_1') &= k\theta \\ E(m_2') &= k\theta(1-\theta) + k^2\theta^2 \end{aligned}$$

Removing expected value signs and adding hats gives us

$$\begin{aligned} m_1' &= \widehat{K}\widehat{\Theta} \\ m_2' &= \widehat{K}\widehat{\Theta}(1-\widehat{\Theta}) + \widehat{K}^2\widehat{\Theta}^2 \end{aligned}$$

We then solve two equations in two unknowns to get

$$\begin{aligned} \widehat{\Theta} &= m_1' - \frac{m_2'}{m_1'} + 1 \\ \widehat{K} &= \frac{m_1'^2}{m_1'^2 + m_1' - m_2'} \end{aligned}$$

Are these estimates unbiased? Good luck trying to compute the expected values! But they are consistent, provided $k\theta \neq 0$. Here is how you would prove it for $\widehat{\Theta}_n$.

By the Law of Large numbers, $m_1' \xrightarrow{P} k\theta$ and $m_2' \xrightarrow{P} k\theta(1-\theta) + k^2\theta^2$. The function $g(x,y) = x - y/x + 1$ is continuous except when $x = 0$, and so by the Continuous Mapping Theorem,

$$\begin{aligned} g(m_1', m_2') &\xrightarrow{P} g(k\theta, k\theta(1-\theta) + k^2\theta^2) \\ &= k\theta - \frac{k\theta(1-\theta+k\theta)}{k\theta} + 1 \\ &= k\theta - 1 + \theta - k\theta + 1 \\ &= \theta, \end{aligned}$$

showing that $\widehat{\Theta}_n$ is consistent for θ, as long as $k \neq 0$ and $\theta \neq 0$.

A Regression Example Let $Y_i = \beta x_i + \epsilon_i$, for $i = 1, \ldots, n$, where

$x_1, \ldots, x_n$ are fixed, known constants

$\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed random variables with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$.

β and σ^2 are unknown parameters.

Let's find a Method of Moments estimator for β.

$$\begin{aligned} E(\overline{Y}_n) &= E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\ &= E\left(\frac{1}{n}\sum_{i=1}^{n}(\beta x_i + \epsilon_i)\right) \\ &= E(\beta\overline{x}_n + \overline{\epsilon}_n) \\ &= \beta\overline{x}_n \end{aligned}$$

So set $\overline{Y}_n = \widehat{\beta}_n \overline{x}_n$, and get

$$\widehat{\beta}_n = \frac{\overline{Y}_n}{\overline{x}_n}.$$

Is it unbiased? Yes, immediately. Consistent? *Be very careful here!* You cannot simply invoke the Law of Large Numbers, because the Law of Large Numbers was proved using Chebyshev's inequality, and the proof of Chebyshev's inequality assumed that $X_1, \ldots, X_n$ all had the same mean μ and the same variance σ^2. It does not apply here, because $E(Y_i)$ depends on i. If you use it directly in a case like this, you might get a point or two out of charity, but you will lose a lot of marks. Even Theorem 10.3 does not apply, because the proof in the book uses Chebyshev's inequality.

On the other hand, Theorem 10.3b is just a theorem about sequences of random variables. The proof uses Markov's inequality, but Markov's inequality always holds. All we need to apply the theorem is the existence of $Var(\widehat{\beta}_n)$ for n sufficiently large. Here we go.

We know $\widehat{\beta}_n$ is unbiased, so it's asymptotically unbiased. We just have to check to see if $Var(\widehat{\beta}_n) \to 0$.

$$\begin{aligned}
Var(\widehat{\beta}_n) &= Var\left(\frac{\overline{Y}_n}{\overline{x}_n}\right) \\
&= \frac{1}{\overline{x}_n^2} Var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\
&= \frac{1}{n^2\overline{x}_n^2} Var\left(\sum_{i=1}^{n} \beta x_i + \epsilon_i\right) \\
&\overset{\text{ind}}{=} \frac{1}{n^2\overline{x}_n^2} \sum_{i=1}^{n} Var(\beta x_i + \epsilon_i) \\
&= \frac{1}{n^2\overline{x}_n^2} \sum_{i=1}^{n} \sigma^2 \\
&= \frac{1}{n^2\overline{x}_n^2} n\sigma^2 \\
&= \frac{\sigma^2}{n\overline{x}_n^2}
\end{aligned}$$

Now we can see the condition for $\widehat{\beta}_n$ to be consistent: $\frac{1}{n\overline{x}_n^2} \to 0$ as $n \to \infty$. A convenient *sufficient* condition is for $\overline{x}_n \to \overline{x}_\infty \neq 0$, but it's more than we really need. The sequence of constants $\overline{x}_n$ can actually converge to zero, as long as it does not go there too fast.

Modified Central Limit Theorem There are many versions of the Central Limit Theorem. This one just says that the version we proved in class still holds when the parameter σ is replaced by any consistent estimator. That is, let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\widehat{\sigma}} \xrightarrow{d} Z \sim N(0,1),$$

where $\widehat{\sigma}$ is *any* consistent estimator of σ.

Hypothesis testing

Model We will write the model as $\mathbf{X} \sim P_\theta$, $\theta \in \Theta$.

- The quantity $\mathbf{X}$ represents the collection of all *observable* random variables in the model. Frequently, $\mathbf{X} = X_1, \ldots, X_n$, but not always. For example,
  - Suppose we have two independent random samples, $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$. Then $\mathbf{X} = X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$.
  - In the simple regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $\mathbf{X} = Y_1, \ldots, Y_n$. The independent variables $x_1, \ldots, x_n$ are not part of $\mathbf{X}$ because they are known, fixed constants. The error terms $\epsilon_1, \ldots, \epsilon_n$ are random, but they are not observable; so, they are not part of $\mathbf{X}$ either.
- P_θ is the joint probability distribution of the random variables in $\mathbf{X}$. It is equivalent to the likelihood. It depends on the parameter θ, which could be a vector. For example, if $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$.
- The probability distribution P_θ can be a function of various known constants as well as unknown parameters. In the regression example, $x_1, \ldots, x_n$ are part of P_θ.

Θ is called the *parameter space*. It is the set of all permissible values for θ. That is, P_θ is a probability distribution if and only if $\theta \in \Theta$. In the case of a normal distribution, $\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$.

We will denote the *sample space* by $\mathcal{X}$. The random vector $\mathbf{X}$ takes values in $\mathcal{X}$, and the joint cumulative distribution of $\mathbf{X}$ is defined on all of $\mathcal{X}$. For example, let $\mathbf{X} = (X_1.X_2)$, where X_1 and X_2 are independent Bernoulli random variables. The sample space $\mathcal{X}$ is the entire plane $\Re^2$. The probability is *zero* that $\mathbf{X}$ falls into a cirle of unit radius, centered at (-14,-22). But that set is still part of the sample space.

In this course, $\mathcal{X}$ will always be the set of real numbers of dimension m, where m is the length of $\mathbf{X}$. That's because all our random variables are real-valued. If our random variables were more exotic (for example, they could be complex-valued), then $\mathcal{X}$ would be more exotic too. But we don't need to go there.

We will denote the *support* of the (joint) distribution P_θ by $\mathcal{S}$, defined as follows. Let

$$\mathcal{S} = \{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}; \theta) > 0\},$$

where $f(\mathbf{x}; \theta)$ is the joint density or distribution (probability mass function) of $\mathbf{X}$. Note that

- The support is contained in the sample space: $\mathcal{S} \subseteq \mathcal{X}$.
- The support $\mathcal{S}$ may depend upon the parameter, but the sample space $\mathcal{X}$ does not.
- As the sample size increases, $\mathcal{X}$ and $\mathcal{S}$ get bigger, while Θ stays the same size.

For example, let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Binomial}(k, \theta)$, where k and θ are both unknown. We have

- $\Theta = \{(k, \theta) : k = 0, 1, \ldots \text{ and } 0 < \theta < 1\}$.
- $\mathcal{X} = \Re^n = \{(x_1, \ldots, x_n) : \infty < x_i < \infty, i = 1, \ldots, n\}$
- $\mathcal{S} = \{(x_1, \ldots, x_n) : x_i = 0, \ldots, k \text{ for } i = 1, \ldots, n\}$.

Notice how the support depends upon the parameter – specifically upon the k part – while the sample space does not.

When specifying Θ and $\mathcal{S}$ for a particular problem, please do so in terms of the technical details of the probability distributions you are using, not in terms of the real-life phenomena being modelled. For example, we might adopt the normal distribution as a model for the number we get when we weigh a randomly selected pig. It's a good model, but of course it's not exactly right. Weights (and expected weights) cannot be negative, but $\mathcal{S}$ still includes negative values of the X_i, and Θ includes negative values of μ.

The General Idea of classical hypothesis testing is this. First, we tentatively adopt an assumption about the value of θ. This assumption is called the "null" hypothesis, because it is supposed to capture the idea that nothing interesting is going on. For example in a blind taste test, the null hypothesis might be that expressed preference for one brand of cola over another is like tossing a fair coin. The model would be $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Bernoulli}(\theta)$, and the null hypothesis is $H_0 : \theta = \frac{1}{2}$.

If the data we observe are *very* unlikely given the assumption of the null hypothesis (like, 90% of the sample prefer Brand A), then we *reject* H_0 in favour of an alternative hypothesis H_1, which is set up to be the positive conclusion of the investigation. For example, $H_1 : \theta \neq \frac{1}{2}$, or $H_1 : \theta > \frac{1}{2}$.

Formal Statement of Null and Alternative Hypotheses We will write

$$H_0 : \theta \in \Theta_0 \text{ v.s. } H_1 : \theta \in \Theta_1,$$
$$\Theta_0 \subset \Theta,\ \Theta_1 \subset \Theta,\ \text{and } \Theta_0 \cap \Theta_1 = \emptyset.$$

It is nice when $\Theta_0 \cup \Theta_1 = \Theta$, but this need not be the case. Recall from the text and lecture that statistical hypotheses can be either *simple* or *composite*. When testing a simple hull hypothesis against a simple alternative, both Θ_0 and Θ_1 are singleton sets consisting on one element each. You definitely do *not* have $\Theta_0 \cup \Theta_1 = \Theta$, except maybe in very strange and artificial cases.

Why would anyone want to leave out part of the parameter space, thus effectively assuming that the parameter cannot *possibly* be in the missing part? I can think of two reasons.

- The classical methods we're studying are formal tools for deciding between two alternatives: $\theta \in \Theta_0$ and $\theta \in \Theta_1$. Sometimes, parts of the parameter space are just irrelevant. Suppose I'm giving patients a test for the level of some blood chemical, and I'm doing it two ways, a cheap way that is relatively inaccurate, and an expensive way that is more accurate. Let X_i represent the cheap measurement and

Y_i represent the expensive measurement for patient i, and denote $Cov(X_i, Y_i)$ by κ, which is the same for $i = 1, \ldots, n$. I am very interested in testing $H_0 : \kappa = 0$ versus $H_1 : \kappa > 0$. The possibility $\kappa < 0$ would mean that on average, high values of X go with low values of Y and vice versa. It's not even worth considering. Similarly, this is where you can eliminate the possibility that the expected weight of your pigs is negative.

- You can also set up simple versus simple hypotheses as a stepping stone to other, more practical comparisons. Applications of the Neyman-Pearson lemma usually fall into this category. More later.

Critical Regions We need a rule for deciding between H_0 and H_1. The decision should be based on the data $\mathbf{X}$, while the *rule* should be something we can set up in advance, before we see the data. Accordingly, we divide the sample space into two disjoint regions, C and C^c. $\mathcal{X} = C \cup C^c$.

The set C is called a *critical region*. It is also called a *rejection region*; the two terms mean the same thing. If $\mathbf{X} \in C$, then we reject H_0 and conclude $\theta \in \Theta_1$. If $\mathbf{X} \notin C$, then we accept H_0 and conclude $\theta \in \Theta_0$.

Several comments are in order.

- We are living in a world of decision theory here. We are deciding between $\theta \in \Theta_0$ and $\theta \in \Theta_1$. These are the only two alternatives, and we have to pick one. [1]

- Usually, part of the support is in C, and part is in C^c. Parts of C and parts of C^c may have probability zero, but that does not present any problem. If the support does not depend on the parameter, then the decision rule is not affected. $P_\theta(\mathbf{X} \in C) = P_\theta(\mathbf{X} \in C \cap \mathcal{S})$, regardless of what θ is, and the same applies to C^c.

  On the other hand, if the support *does* depend on θ, it is very nice when C contains the regions of $\mathcal{X}$ that have zero probability under H_0 — because if $\mathbf{X}$ falls into there, then the null hypothesis is definitely wrong, and the decision to reject it is correct, with probability one.

- A critical region is equivalent to a test. Sometimes we call C a critical region, and sometimes we call it a test. There should be no confusion.

[1] In some classes (like STA220, STA221 and STA442), students are told that if we reject H_0 then we accept H_1, but if we do *not* reject H_0, then we conclude nothing. We *never* accept the null hypothesis, they are told. This point of view is due to R. A. Fisher (Mr. F distribution) who came up with the concept of hypothesis testing more or less as we know it. As usual with intellectual pioneers, his ideas were brilliant but not every detail was precisely worked out. Neyman and Pearson came along a bit later and cleaned up Fisher's method, putting it on a firm decision-theoretic basis.

During their lifetimes, Fisher fought bitterly with Neyman and Pearson. To Neyman and Pearson, Fisher was creative but mathematically unsophisticated. To Fisher, Neyman and Pearson were good mathematicians, but they were missing the point, because science does not proceed by simple yes or no decisions.

Today, Neyman-Pearson theory usually dominates in theoretical research and theoretical courses, while Fisher's approach is more helpful in applications and applied courses. STA 261 is a theoretical course, and so we're in Neyman-Pearson mode for the moment.

- Almost always, critical regions are defined in terms of *test statistics*. For example, $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, and we want to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. Consider the critical region $C = \{\mathbf{x} \in \mathcal{X} : \left|\frac{\overline{x}-\mu_0}{s/\sqrt{n}}\right| > t_{\alpha/2,n-1}\}$. Clearly $\mathbf{X} \in C$ if and only if $|T(\mathbf{X})| > t_{\alpha/2,n-1}$, where $T = \frac{\overline{X}-\mu_0}{S/\sqrt{n}}$. The test statistic is T.

- Statistics cannot depend upon unknown parameters. And, whether you write a critical region in terms of a statistic or not, neither can a critical region. Again, *a critical region cannot be defined in terms of any unknown parameters*. It only makes sense. A critical region is equivalent to a decision rule, and you cannot make decisions based on something you do not know.

- For almost any model and null and alternative hypothesis, a huge number of tests (critical regions) are possible. Some must be better than others. Better in what sense? This is the next point to consider.

Type I and Type II Error When we perform a test — that is, when we observe a value of $\mathbf{X}$ and check whether it falls into C — four things are possible.

1. $\theta \in \Theta_0$ but we mistakenly conclude $\theta \in \Theta_1$.
2. $\theta \in \Theta_1$ but we mistakenly conclude $\theta \in \Theta_0$.
3. $\theta \in \Theta_0$ and we correctly conclude $\theta \in \Theta_0$.
4. $\theta \in \Theta_1$ and we correctly conclude $\theta \in \Theta_1$.

If $\theta \in \Theta_0$ and it happens that $\mathbf{X} \in C$, we falsely conclude $\theta \in \Theta_1$. This is called a *Type I Error*. If $\theta \in \Theta_1$ and it happens that $\mathbf{X} \notin C$, we falsely conclude $\theta \in \Theta_0$. This is called a *Type II Error*. We would like to choose C so as to minimize the probability of both kinds of error, simultaneously. Unfortunately, this is not possible; you can't do both.

Consider the awful strategy of *never* rejecting H_0, regardless of the data. The probability of Type II error may be high, but the probability of Type I error is zero. For minimizing Type I error, you cannot beat this "test." At the other extreme, you could *always* reject H_0, regardless of the data. The probability of Type I error may be high, but the probability of Type II error is zero. We need a principled tradeoff between the two types of error.

Here's the standard answer. Restrict your attention to critical regions C such that for $\theta \in \Theta_0$, $P_\theta\{\mathbf{X} \in C\}$ is held to a fixed, low level. Then seek a critical region with $P_\theta\{\mathbf{X} \in C\}$ large for $\theta \in \Theta_1$.

Size of a test The *size* of a test is the maximum probability of making a Type I Error. In symbols, it is

$$\alpha = \max_{\theta \in \Theta_0} P_\theta\{\mathbf{X} \in C\}$$

This is a standard definition, but it does not seem to be in our text. Our authors treat the probability of a Type I error as a function of θ, writing it as $\alpha(\theta)$ for $\theta \in \Theta_0$. Mostly they avoid composite null hypotheses. We will not. The probability of a Type II error is also definitely a function of θ. It is written $\beta(\theta)$.

Power The power of a test is the probability of correctly rejecting the null hypothesis. That is, it is the probability of detecting something (like a difference between expected values, perhaps) that is really present. It too is a function of θ: Power $= 1 - \beta(\theta)$ for $\theta \in \Theta_1$.

Our text defines the overall *power function* $\pi(\theta)$ as $\alpha(\theta)$ for $\theta \in \Theta_0$ and $1 - \beta(\theta)$ for $\theta \in \Theta_1$. In our notation,

$$\pi(\theta) = P_\theta\{\mathbf{X} \in C\} = \alpha(\theta)I(\theta \in \Theta_0) + (1 - \beta(\theta))I(\theta \in \Theta_1).$$

Clearly, power (statistical power, anyway) is a good thing. If C and D are two critical regions of size α with $P_\theta\{\mathbf{X} \in C\} > P_\theta\{\mathbf{X} \in D\}$, we say C is *more powerful* than D — for that particular value of θ. If C is more powerful than any other size α critical region, it is said to be the *most powerful* test. This is great, but it is still specific to that θ. If C is most powerful for *all* $\theta \in \Theta_1$, the test is said to be *uniformly most powerful* of size α. You can't do any better than that. It's a wonder that any such tests exist. Thank you, Mr. Neyman. Thank you, Mr. Pearson.

The Neyman-Pearson Lemma The key to deriving uniformly most powerful tests (when they exist) is this lemma, which gives us the most powerful test of a simple null hypothesis versus a simple alternative.

Without losing any generality, we can write $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$. Write the likelihood as $L(\mathbf{x}; \theta)$. The Neyman-Pearson Lemma says that the most powerful size α test is given by

$$C = \{\mathbf{x} \in \mathcal{X} : L(\mathbf{x}; \theta_0) \leq k\, L(\mathbf{x}; \theta_1)\}.$$

The constant k is general, and is chosen to make the test have size α. That is, $P_{\theta_0}\{\mathbf{X} \in C\} = \alpha$, where it is recognized that for discrete distributions, we may not be able to have any exact value of α we want[2]. Also, because any critical region consists of points in $\mathcal{X}$, the "$\in \mathcal{X}$" part in the definition of C is redundant, and may be dropped in long calculations.

When we say that C is most powerful of size α, we mean precisely the following. Let D be *any* other critical region with $P_{\theta_0}\{\mathbf{X} \in D\} \leq \alpha$. Then $P_{\theta_1}\{\mathbf{X} \in C\} \geq P_{\theta_1}\{\mathbf{X} \in D\}$. The proof given in lecture (but not the one in the book) carefully includes the case where the case where $P_{\theta_0}\{\mathbf{X} \in D\}$ is strictly less than α.

Our book writes the likelihoods in a *ratio*, which is often useful in calculations, and also nice because it makes intuitive sense. If the ratio $\frac{L(\mathbf{x};\theta_0)}{L(\mathbf{x};\theta_1)}$ is small, it means that the observed data are relatively less probable given $\theta = \theta_0$ than they are given $\theta = \theta_1$. If the relative likelihood is small enough, then surely $H_0 : \theta = \theta_0$ should be rejected. Still, we're not writing it as a ratio here, because sometimes the likelihood equals zero when the support of the distribution depends on the parameter – and we want to avoid division by zero.

Sample Questions and answers Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with mean μ and variance one. We will test $H_0 : \mu \leq \mu_0$ versus $H_0 : \mu > \mu_0$ with the test statistic $T = \sqrt{n}(\overline{X} - \mu_0)$, rejecting H_0 if $T > z_\alpha$.

[2]Well, you can if you use a *randomized* test, but randomized tests are so unacceptable in applications that they are not discussed here.

Question 1 *What is the parameter space Θ?*

Answer: The set of real numbers. That is, $\{\mu : -\infty < \mu < \infty\}$.

Question 2 *What is Θ_0?*

Answer: $(-\infty, \mu_0]$

Question 3 *What is Θ_1?*

Answer: (μ_0, ∞)

Question 4 *What is the sample space $\mathcal{X}$?*

Answer: $\Re^n$, not just $\Re$.

Question 5 *What is the support $\mathcal{S}$?*

Answer: Again, $\Re^n$. The normal density does not require an indicator for the support.

Question 6 *What is the critical region C?*

Answer: $\{\mathbf{x} \in \Re^n : \sqrt{n}(\overline{x} - \mu_0) > z_\alpha\}$.

Question 7 *Find the power function $\pi(\mu)$.*

Answer: We know $\overline{X} \sim N(\mu, 1/n)$, and of course μ need not equal μ_0. So,

$$\begin{aligned}
P_\mu\{\mathbf{X} \in C\} &= P_\mu\{\sqrt{n}(\overline{X} - \mu_0) > z_\alpha\} \\
&= P_\mu\{\overline{X} > \frac{z_\alpha}{\sqrt{n}} + \mu_0\} \\
&= P_\mu\{\sqrt{n}(\overline{X} - \mu) > \sqrt{n}(\frac{z_\alpha}{\sqrt{n}} + \mu_0 - \mu)\} \\
&= P_\mu\{\sqrt{n}(\overline{X} - \mu) > z_\alpha + \sqrt{n}(\mu_0 - \mu)\} \\
&= 1 - F_Z[z_\alpha + \sqrt{n}(\mu_0 - \mu)],
\end{aligned}$$

where F_Z is the cumulative distribution function of a standard normal.

Question 8 *Find the size of the test. Hint: H_0 is composite.*

Answer: Because the null hypothesis is composite, we must *maximize* $\pi(\theta)$ over Θ_0.

$$\begin{aligned}
\frac{d}{d\mu} P_\mu\{\mathbf{X} \in C\} &= \frac{d}{d\mu}\left(1 - F_Z[z_\alpha + \sqrt{n}(\mu_0 - \mu)]\right) \\
&= (-1) f_z(z_\alpha + \sqrt{n}(\mu_0 - \mu))(-\sqrt{n}) \; > \; 0,
\end{aligned}$$

so the function is increasing. It attains its maximum on the right boundary of Θ_0, where $\mu = \mu_0$. At that point,

$$1 - F_Z[z_\alpha + \sqrt{n}(\mu_0 - \mu)] = 1 - F_Z[z_\alpha] = 1 - (1 - \alpha) = \alpha.$$

So, the size of the test is α. Note: If the null hypothesis is composite and you are asked to find the size of the test, you **must** maximize the power function over Θ_0 to get full marks. Almost always, the maximum occurs at the point in Θ_0 that is closest to Θ_1, but you have to show it, not just guess.

Question 9 *Suppose $\mu_0 = 0$ and $\alpha = 0.05$. If $n = 4$, what is the power at $\mu = 1$?*

Answer:

$$\begin{aligned} 1 - F_Z[z_\alpha + \sqrt{n}(\mu_0 - \mu)] &= 1 - F_Z[1.645 + 2(0 - 1)] = 1 - F_Z[-0.355] \\ &= 0.5 + \frac{1}{2}(0.1368 + 0.1406) = 0.6387 \end{aligned}$$

Question 10 *Again, suppose $\mu_0 = 0$ and $\alpha = 0.05$. What sample size is required so that if the true value of μ is one half, the power (probability of rejecting H_0) will be at least 0.99?*

Answer: We want

$$\begin{aligned} 0.99 &= 1 - F_Z[1.645 + \sqrt{n}(0 - \frac{1}{2})] \\ &= 1 - F_Z[1.645 - \frac{\sqrt{n}}{2}], \end{aligned}$$

so we set

$$1.645 - \frac{\sqrt{n}}{2} = z_{0.99} = -z_{0.01} = -2.326$$

and solve for n to get $n = 63.08$. The next highest integer is $n = 64$, which makes the probability of rejecting H_0 a bit more than 0.99.

I got the value 2.326 from the last row of the t table. By the way, this is the standard method for selecting sample size when the purpose of a study is to test a hypothesis. You pick a value of $\theta \in \Theta_1$ that is is scientifically reasonable, even modest. Then you find the sample size that makes the probability of rejecting H_0 (and thus concluding H_1) comfortably high for that parameter value. This procedure is quite general, and is usually called *statistical power analysis* in applied statistics.

Question 11 *Prove that this test is uniformly most powerful of size α for testing $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$.*

Let D be another critical region of size α for testing $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$. We seek to show $P_\mu\{\mathbf{X} \in D\} \leq P_\mu\{\mathbf{X} \in C\}$ for all $\mu > \mu_0$.

First, consider the simple null hypothesis $H_0 : \mu = \mu_0$ versus the simple alternative $H_1 : \mu = \mu_1$, where $\mu_1 > \mu_0$. By the Neyman-Pearson lemma, the most powerful test of H_0 versus H_1 is given by

$$\begin{aligned} C &= \{\mathbf{x} \in \mathcal{X} : L(\mathbf{x}; \mu_0) \leq k\, L(\mathbf{x}; \mu_1)\} \\ &= \{\mathbf{x} : \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i - \mu_0)^2}{2}} \leq k \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i - \mu_1)^2}{2}}\} \\ &= \{\mathbf{x} : \exp[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu_0)^2] \leq k\, \exp[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu_1)^2]\} \\ &= \{\mathbf{x} : \frac{\exp[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu_0)^2]}{\exp[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu_1)^2]} \leq k\} \end{aligned}$$

$$
\begin{aligned}
&= \{\mathbf{x} : \exp \frac{1}{2} \sum_{i=1}^{n} [(x_i - \mu_1)^2 - (x_i - \mu_0)^2] \leq k\} \\
&= \{\mathbf{x} : \sum_{i=1}^{n} [(x_i - \mu_1)^2 - (x_i - \mu_0)^2] \leq 2 \ln k = k_1\} \\
&= \{\mathbf{x} : \sum_{i=1}^{n} [x_i^2 - 2x_i\mu_1 + \mu_1^2 - x_i^2 + 2x_i\mu_0 - \mu_0^2] \leq k_1\} \\
&= \{\mathbf{x} : 2(\mu_0 - \mu_1) \sum_{i=1}^{n} x_i + n(\mu_1^2 - \mu_0^2) \leq k_1\} \\
&= \{\mathbf{x} : 2(\mu_0 - \mu_1) \sum_{i=1}^{n} x_i \leq k_2 = k_1 - n(\mu_1^2 - \mu_0^2)\} \\
&= \{\mathbf{x} : (\mu_0 - \mu_1) \sum_{i=1}^{n} x_i \leq k_3 = \frac{1}{2} k_2\} \\
&= \{\mathbf{x} : \sum_{i=1}^{n} x_i \geq k_4 = \frac{k_3}{\mu_0 - \mu_1} k_2\},
\end{aligned}
$$

where the direction of the inequality changed because we divided by $\mu_0 - \mu_1$, a negative number. Notice that although the constant on the right side keeps changing, the set of $\mathbf{x}$, and hence the critical region we are simplifying, remains the same. This is typical of calculations related to the Neyman-Pearson Lemma. The rule is that you can absorb all kinds of constants into k, but **never** x values or functions of $\mathbf{x}$. If you absorb any part of $\mathbf{x}$ into the constant, the answer is wrong. On the other hand, you will not lose marks for failure to write each constant as an explicit function of the last one; I are doing it here just for clarity.

Note also that this critical region does not depend on any unknown parameters; μ_0 is a specific number specified by H_0, and μ_1 is a specific number specified by H_1. Now we continue.

$$
\begin{aligned}
C &= \{\mathbf{x} : \sum_{i=1}^{n} x_i \geq k_4\} \\
&= \{\mathbf{x} : \overline{x} \geq k_5 = k_4/n\} \\
&= \{\mathbf{x} : \sqrt{n}(\overline{x} - \mu_0) \geq k_6 = \sqrt{n}(k_5 - \mu_0)\}.
\end{aligned}
$$

Now choose $k_6 = z_\alpha$, obtaining a size α test. You could reverse all the steps and solve for the original k, but why bother? Nobody ever does it.

This test is most powerful for testing the simple null hypothesis $H_0 : \mu = \mu_0$ versus the simple alternative $H_1 : \mu = \mu_1$. But notice that the final form of the critical region depends upon the value μ_1 only in one subtle way. Because $\mu_1 > \mu_0$, the direction of the inequality reversed at one point in the calculation; that's why it's pointing right instead of left.

Hence, the conclusion applies to *all* $\mu_1 > \mu_0$. We conclude that the test C is *uniformly* most powerful for testing the simple null hypothesis $H_0 : \mu = \mu_0$ versus the composite alternative $H_1 : \mu > \mu_0$. Now we will extend it to the *composite* null hypothesis $H_0 : \mu \leq \mu_0$, which was our goal from the beginning.

Since we have already shown that the power function is increasing, we know that the test C is size α for testing the composite null hypothesis $H_0 : \mu \leq \mu_0$. Now we will show it is uniformly most powerful of size α for testing the composite null hypothesis.

Let D be another critical region of size α for testing $H_0 : \mu \leq \mu_0$. Size α means $P_\mu\{\mathbf{X} \in D\} \leq \alpha$ for all $\mu \in (-\infty, \mu_0]$. Since μ_0 is definitely in this set, we have $P_{\mu_0}\{\mathbf{X} \in D\} \leq \alpha$. Thus D is also a size α test of the *simple* null hypothesis $H_0 : \mu = \mu_0$ versus the composite alternative $H_1 : \mu > \mu_0$. But we have shown that C is uniformly most powerful for this situation; so, $P_\mu\{\mathbf{X} \in D\} \leq P_\mu\{\mathbf{X} \in C\}$ for all $\mu > \mu_0$. ■

Likelihood Ratio Tests

Reconciling the notations In the textbook's section on likelihood ratio tests, they introduce notation for the parameter space etc. that is different from the notation in this handout. Here is a key for translating between the two notations. You may use either one.

- **Parameter space**: $\Omega = \Theta$. The book uses Ω to denote the parameter space, while in lecture and this handout, the parameter space is Θ. The book writes $\theta \in \Omega$. We write $\theta \in \Theta$.
- **Null Hypothesis**: $\omega = \Theta_0$. The book writes $H_0 : \theta \in \omega$. We write $H_0 : \theta \in \Theta_0$.
- **Alternative Hypothesis**: $\omega' = \Theta_1$. The book writes $H_1 : \theta \in \omega'$. We write $H_1 : \theta \in \Theta_1$. In the book's notion, $\omega \cup \omega' = \Omega$. Thus, their null and alternative hypotheses cover the entire parameter space — which is appealing, but their notation does not apply to the Neyman-Pearson section.
- **Maximum Likelihood Estimates**: In both the text and lectures, $\widehat{\theta}$ has been used to denote an estimate of θ, including maximum likelihood estimates. A Maximum Likelihood Estimate, of course, is the point that maximizes the likelihood over the entire parameter space. But in this section of the text,

$\widehat{\theta}$ is the point that maximizes $L(\theta, \mathbf{x})$ over ω, and
$\widehat{\widehat{\theta}}$ is the point that maximizes $L(\theta, \mathbf{x})$ over the whole parameter space Ω.

I suppose the idea is that $\widehat{\widehat{\theta}}$ maximizes over two susets of the parameter space, ω and ω', so it deserves two hats. But it's unfortunate when an important symbol like $\widehat{\theta}$ is used to represent two different quantities that could easily be confused.

So I will continue to use $\widehat{\theta}$ for the unrestricted MLE, and I will use $\widetilde{\theta}$ to denote the restricted MLE. That is,

$$\max_{\theta \in \Theta} L(\theta, \mathbf{x}) = L(\widehat{\theta}, \mathbf{x}) \text{ and } \max_{\theta \in \Theta_0} L(\theta, \mathbf{x}) = L(\widetilde{\theta}, \mathbf{x})$$

See? The hat on the restricted MLE is crushed, broken. It can never be as impressive as the unrestricted MLE. $\widehat{\theta}$ is the location of the tallest mountain in the world. $\widetilde{\theta}$ is the location of the tallest mountain in North America.

Of course I like my notation more, but it does not matter which one you use. The end product will be the same.