BSTA 552: Mathematical Statistics II – Class Notes

Jessica Minnier

2019-10-02

1 BSTA 552

In BSTA 551 you reviewed principles of probability, explored methods for reduction of data and learned how to find point estimators. In this class, you will use those point estimators to test hypotheses about your data and obtain inference about the underlying population parameters by constructing interval estimators. We will then discuss asymptotic properties of these estimators and relate these properties to common statistical tests. Statisticians perform hypothesis testing and construct confidence intervals every day. This class will build the foundation upon which those tests obtain their validity.

The main material of BSTA 552 will continue as in 551 to be derived from Casella and Berger's "Statistical Inference" and the authors' rigorous treatment of the theory of these topics.

C&B will define the backbone of the theory, and will be an indispensable resource to you when studying for comprehensive exams, as well as in your future careers.

2 Hypothesis Testing (C&B 8)

2.1 Definitions and Intro (8.1)

Definition 8.1.1 A hypothesis is a statement about a population parameter.

Definition 8.1.2 In a hypothesis testing problem, we have two competing hypotheses, the null H_0, and alternative, H_1 (or H_a), hypotheses.

The general form of the pair of hypotheses is

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \in \Theta_0^c,$$

where Θ_0 is some subset of the parameter space and Θ_0^C is the complement of Θ_0. We will often write these more simply, for example:

$$H_0 : \theta = \theta_0 \text{ and } H_1 : \theta \neq \theta_0 \text{ or } H_0 : \theta \leq \theta_0 \text{ and } H_1 : \theta > \theta_0$$

Definition 8.1.3 A hypothesis testing procedure or hypothesis test is a rule that specifies:

1. The **rejection (critical) region**: for which sample values H_0 is rejected and H_1 is accepted as true.
2. The **acceptance region**: for which sample values the decision is made to accept H_0 as true.

How to describe a hypothesis test? A rule which says for which sample values we will reject the null hypothesis and for which values we will "not reject" the null hypothesis.

In the textbook C&B mention the controversy over the use of the phrase "accept the null hypothesis." For the most part, we won't use that terminology in making conclusions, although we will use it for discussion of types of errors in hypothesis testing.

Notes

- Typically, a hypothesis test is specified in terms of a *test statistic* $W(X_1, \ldots, X_n) = W(\mathbf{X})$, a function of the sample.
- Examples of a *test statistic*: $W(X_1, \ldots, X_n) = n^{-1} \sum_{i=1}^{n} X_i$ or $W(X_1, \ldots, X_n) = X_2$ or $W(X_1, \ldots, X_n) = X_{(1)}$
- Like point estimators, tests *must be evaluated* before we can determine whether they provide useful inference. But first we must describe methods of finding tests.
- Definition of rejection and acceptance regions involves using evidence against or for the null. The evidence comes from our sample data!

2.2 Methods of Finding Tests (8.2)

There are various approaches for finding hypothesis tests.

We will start with likelihood ratio tests, which have connections to maximum likelihood estimation.

2.2.1 Likelihood Ratio Tests (LRTs)

- **Goal**: compare the likelihood of the sample data under the null and the alternative hypotheses.
- If the likelihood of the sample data under the alternative hypothesis is significantly greater than under the null hypothesis then we will reject the null hypothesis in favor of the alternative hypothesis.
- The rule is typically defined in terms of a *test statistic* involving. . . you guessed it, likelihoods!

Recall that if the join pdf (pmf) for data $X_1, \dots, X_n$ is denoted by $f(\mathbf{x}|\theta)$, the *likelihood function* is:

$$\mathcal{L}(\theta|\mathbf{x}) = \mathcal{L}(\theta|x_1, \dots, x_n) = f(\mathbf{x}|\theta),$$

where θ could be a scalar or a vector.

Let Θ denote the full parameter space.

Definition 8.2.1 LRT The *likelihood ratio test statistic* for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ is[1]

$$\lambda(\mathbf{x}) = \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta|\mathbf{x})}{\sup_{\theta \in \Theta} \mathcal{L}(\theta|\mathbf{x})}.$$

A *likelihood ratio test (LRT)* is any test that has a rejection region of the form

$$\{\mathbf{x} : \lambda(\mathbf{x}) \leq c\},$$

where c is any number satisfying $0 \leq c \leq 1$.

Note that if the likelihood of the sample is greater for some value of the parameter, θ, under the alternative compared to all values of the parameter under the null hypothesis, the denominator of the LRT statistic will be greater than the numerator, and hence the LRT statistic will be less than 1. In this case, we think of the alternative as more likely than the null hypothesis. (small/large)

Recall that MLEs were defined as maximizing the likelihood functions. If we maximize the likelihood in the restricted null hypothesis space and maximize the likelihood over the entire parameter space we have found which θ estimates give us the top and bottom pieces of the LRT fraction. Thus we can express the LRT statistic using MLE's as follows:

$$\boxed{\lambda(\mathbf{x}) = \frac{\mathcal{L}(\widehat{\theta}_0|\mathbf{x})}{\mathcal{L}(\widehat{\theta}|\mathbf{x})}}$$

where $\widehat{\theta}_0$ denotes a restricted MLE obtained by maximizing the likelihood over a restricted parameter space. Specifically, $\widehat{\theta}_0 = \widehat{\theta}_0(\mathbf{x})$, is the value of $\theta \in \Theta_0$ that maximizes $\mathcal{L}(\theta|\mathbf{x})$.

Example (*Example 8.2.2, Normal LRT*): Let $X_1, \dots, X_n$ be a random sample from a $\mathcal{N}(\theta, 1)$ population. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Here θ_0 is a number fixed by the experimenter prior to the experiment. Since there is only one value of θ specified by H_0, the

[1] sup = supremum, or least upper bound, is similar to maximum but is more general; a supremum of set S is the smallest upper bound of S, where an upper bound is a number B such that $x \leq B$ for all $x \in S$, and if the upper bound is in S then it is the maximum of S.

numerator of $\lambda(\mathbf{x})$ is $\mathcal{L}(\theta_0|\mathbf{x})$. In Example 7.2.5 the (unrestricted) MLE of θ was found to be $\bar{\mathbf{X}}$, the sample mean. Thus the denominator of $\lambda(\mathbf{x})$ is $\mathcal{L}(\bar{\mathbf{x}}|\mathbf{x})$. So the LRT statistic is:

$$\begin{aligned}
\lambda(\mathbf{x}) &= \frac{(2\pi)^{-n/2}\exp[-\sum_{i=1}^n (x_i-\theta_0)^2/2]}{(2\pi)^{-n/2}\exp[-\sum_{i=1}^n (x_i-\bar{\mathbf{x}})^2/2]} \\
&= \exp\left[\left(-\sum_{i=1}^n (x_i-\theta_0)^2 + \sum_{i=1}^n (x_i-\bar{\mathbf{x}})^2\right)/2\right] \\
&= \exp\left[\left(-\sum_{i=1}^n (x_i-\bar{\mathbf{x}}+\bar{\mathbf{x}}-\theta_0)^2 + \sum_{i=1}^n (x_i-\bar{\mathbf{x}})^2\right)/2\right] \\
&= \exp\left[\left(-\sum_{i=1}^n (x_i-\bar{\mathbf{x}})^2 - 2(\bar{\mathbf{x}}-\theta_0)\sum_{i=1}^n (x_i-\bar{\mathbf{x}}) - n(\bar{x}-\theta_0)^2 + \sum_{i=1}^n (x_i-\bar{\mathbf{x}})^2\right)/2\right] \\
&= \exp[-n(\bar{\mathbf{x}}-\theta_0)^2/2]
\end{aligned}$$

For this test statistic to lend itself to become a test procedure, we need to define the rejection and acceptance regions. An LRT is a test that rejects H_0 for small values of $\lambda(\mathbf{x})$. From (8.2.2), the rejection region $\{\mathbf{x} : \lambda(\mathbf{x}) \leq c\}$, can be written as

$$\{\mathbf{x} : |\bar{\mathbf{x}} - \theta_0| \geq \sqrt{-2(\log c)/n}\}$$

As c ranges between 0 and 1, $\sqrt{-2(\log c)/n}$ ranges between 0 and ∞. Thus, the LRTs are just those tests that reject $H_0 : \theta = \theta_0$ if the sample mean differs from the hypothesized value θ_0 by more than a specified amount. (Note this test procedure is not very useful unless we choose a c, more on this later).

Recap: First find the expression for $\lambda(\mathbf{X})$, then we can simplify the rejection region as an expression involving $|\bar{\mathbf{X}} - \theta_0|$ (remember $\bar{\mathbf{X}}$ is a sufficient statistic). This simplification is a recurring theme (and will become theorem 8.2.4.)

Example (*Example 8.2.3, Exponential LRT*)

But suppose the solution to the optimization is not a simple interior point. Let $X_1, \ldots, X_n$ be a random sample from an exponential population with pdf

$$f(x) = \begin{cases} e^{-(x-\theta)}, & x \geq \theta \\ 0, & x < \theta \end{cases}$$

where $-\infty < \theta < \infty$. Note how the support of $\mathbf{x}$ depends on θ. You typically do not know the value of θ and so you must estimate it. The likelihood function is:

$$\mathcal{L}(\theta|\mathbf{x}) = \begin{cases} e^{-\sum_{i=1}^n x_i + n\theta}, & \theta \leq x_{(1)} \\ 0, & \theta > x_{(1)} \end{cases}$$

Now suppose you (the experimenter and the analyst) choose a θ_0 and wish to test:

$$H_0 : \theta \leq \theta_0$$

$$H_1 : \theta > \theta_0$$

We have two cases, $x_{(1)} \leq \theta_0$ or $x_{(1)} > \theta_0$.

We have $\Theta = -\infty < \theta \leq x_{(1)}$ (we know this is the possible range of values for θ, given our model and pdf), we see that if we maximize the likelihood in this parameter space we see the maximum at $x_{(1)}$:

$$\mathcal{L}(\widehat{\theta}_0|\mathbf{x}) = \mathcal{L}(x_{(1)}|\mathbf{x}) = e^{-\sum_{i=1}^{n} x_i + nx_{(1)}}$$

since the function is increasing in θ and $\theta \leq x_{(1)}$. This is the denominator of $\lambda(\mathbf{x})$, the unrestricted maximum of $\mathcal{L}(\theta|\mathbf{x})$.

Now we need the restricted maximum (supremum) in the null hypothesis parameter space for our numerator. We think of our cases:

If $x_{(1)} > \theta_0$, we have $\sup_{\Theta_0} \mathcal{L}(\theta|\mathbf{x}) = e^{-\sum_{i=1}^{n} x_i + n\theta_0}$ as this is the largest θ under H_0 (actually the limit, since we need the supremum).

If $x_{(1)} \leq \theta_0$, we have $\sup_{\Theta_0} \mathcal{L}(\theta|\mathbf{x}) = e^{-\sum_{i=1}^{n} x_i + nx_{(1)}}$ because the model requires $\theta \leq x$.

Hence, we have our LRT:

$$\lambda(\mathbf{x}) = \begin{cases} 1, & x_{(1)} \leq \theta_0 \\ e^{-n(x_{(1)} - \theta_0)}, & x_{(1)} > \theta_0. \end{cases}$$

Again we are left with our LRT simplified in an expression of a sufficient statistic for θ.

We actually could have started this by using the distribution of the sufficient statistic $x_{(1)}$, because of the following theorem.

Theorem 8.2.4 – LRT's and Sufficient Statistics: If $T(\mathbf{X})$ is a sufficient statistic for θ and $\lambda^*(t)$ and $\lambda(\mathbf{x})$ are the LRT statistics based on T and $\mathbf{X}$, respectively, then $\lambda^*(T(\mathbf{x})) = \lambda(\mathbf{x})$ for every $\mathbf{x}$ in the sample space.

Proof: Recall from the Factorization Theorem (Theorem 6.2.6), the pdf or pmf of $\mathbf{X}$ can be written as $f(\mathbf{x}|\theta) = g(T(\mathbf{x})\theta)h(\mathbf{x})$, where $g(t|\theta)$ is the pdf or pmf of T and $h(\mathbf{x})$ does not depend on θ. Thus

$$\begin{aligned}
\lambda(\mathbf{x}) &= \frac{\sup_{\theta\in\Theta_0} \mathcal{L}(\theta|\mathbf{x})}{\sup_{\theta\in\Theta} \mathcal{L}(\theta|\mathbf{x})} \\
&= \frac{\sup_{\theta\in\Theta_0} f(\mathbf{x}|\theta)}{\sup_{\theta\in\Theta} f(\mathbf{x}|\theta)} \\
&= \frac{\sup_{\theta\in\Theta_0} g(T(\mathbf{x})|\theta)h(\mathbf{x})}{\sup_{\theta\in\Theta} g(T(\mathbf{x})|\theta)h(\mathbf{x})} \\
&= \frac{\sup_{\theta\in\Theta_0} g(T(\mathbf{x})|\theta)}{\sup_{\theta\in\Theta} g(T(\mathbf{x})|\theta)} \\
&= \frac{\sup_{\theta\in\Theta_0} \mathcal{L}^*(\theta|T(\mathbf{x}))}{\sup_{\theta\in\Theta} \mathcal{L}^*(\theta|T(\mathbf{x}))} \\
&= \lambda^*(T(\mathbf{x})).
\end{aligned}$$

$\mathcal{L}^*(\theta|T(\mathbf{x}))$ is the likelihood based on T.

So, the likelihood ratio test statistic depends on the data only through T if T is sufficient for θ.

Summary: Sometimes it is easier to reduce by sufficiency first to find the form of the LRT.

Note: Likelihood ratio tests are also useful in situations where there are nuisance parameters, that is, parameters that are present in a model but are not of direct inferential interest. For example, when we have a normally distributed variable and only care to make inference on the true mean but do not know the true variance (or care to know it).

Example (*Example 8.2.6, Normal LRT with unknown variance*)

In the case where we have normally distributed data and want to make inference on the mean, the standard deviation σ becomes a nuisance parameter. Let $X_1, \ldots, X_n$ be a random sample from a $\mathcal{N}(\mu, \sigma^2)$ population. Consider testing $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$. The LRT statistic is

$$\lambda(\mathbf{x}) = \frac{\max_{\mu,\sigma:\mu\leq\mu_0,\sigma^2>0} \mathcal{L}(\theta|\mathbf{x})}{\max_{\mu,\sigma:-\infty<\mu<\infty,\sigma^2>0} \mathcal{L}(\theta|\mathbf{x})} = \frac{L(\widehat{\mu}_0, \widehat{\sigma}_0^2)}{L(\widehat{\mu}, \widehat{\sigma}^2)}$$

If $\widehat{\mu} > \mu_0$

$$\begin{aligned}
\widehat{\mu} = \bar{X},\ \widehat{\sigma}^2 &= \sum_{i=1}^{n}(X_i - \bar{X})/n \\
\widehat{\mu}_0 = \mu_0,\ \widehat{\sigma}_0^2 &= \sum_{i=1}^{n}(X_i - \mu_0)/n
\end{aligned}$$

while if $\widehat{\mu} < \mu_0$

$$\widehat{\mu} = \widehat{\mu}_0 = \bar{X},\, \widehat{\sigma}^2 = \widehat{\sigma}_0^2 = \sum_{i=1}^{n} (X_i - \bar{X})/n$$

so we have

$$\lambda(\mathbf{x}) = \begin{cases} 1, & \widehat{\mu} \leq \mu_0 \\ \frac{L(\widehat{\mu}, \widehat{\sigma}^2)}{L(\widehat{\mu}_0, \widehat{\sigma}_0^2)}, & \widehat{\mu} > \mu_0. \end{cases}$$

Example: Two sample exponential

Suppose that we have two independent random samples: $X_1, \ldots X_n$ are exponential(θ) and $Y_1, \ldots, Y_m$ are exponential(μ). Find the LRT of $H_0 : \theta = \mu$ versus $H_1 : \theta \neq \mu$.

2.3 Review

What do we know about likelihoods?

1. The *likelihood of the sample* is a function of the parameter θ
2. It is defined as the joint probability (discrete sample) or joint density (continuous sample) of the sample data
3. We can then ask: Given the data we've collected we can measure how "likely" was the data generated from the specific pmf/pdf $f(\theta)$ for each possible value of θ?
4. Factorization Theorem: If (and only if) it can be factored into two functions $\mathcal{L}(\theta|\mathbf{X}) = g(T(\theta)|\mathbf{X})h(\mathbf{X})$ then $T(\mathbf{X})$ is a sufficient statistic for θ
5. Under the Likelihood Principle, if two likelihood functions from two sample points $\mathbf{x}$ and $\mathbf{y}$ are proportional (as functions of θ) then the two sample points will give identical inference on θ.
6. We can create a Likelihood Ratio Test if we maximize the likelihood on a restricted null hypothesis parameter space and on the full space and take the ratio
7. We can maximize likelihoods to obtain an MLE estimator $\widehat{\theta}$ which has "nice" properties
   - Invariance property: $t(\widehat{\theta})$ is the MLE of $t(\theta)$
   - Asymptotically (for large-samples) under regularity conditions, $\widehat{\theta}$ is consistent (asymptotically unbiased) and normally distributed (the nicest of distributions) with mean θ and known variance related to the Information

What do we know about hypothesis tests?

1. A hypothesis test involves making a statement about parameter(s) and then using a statistic to compare the observed sample with the theory
2. The test statistic (like an estimator) is a function of the sample measurements.
3. The test statistic (like an estimator) is a random variable.
4. The rejection region specifies the values of the test statistic for which the null hypothesis is to be rejected in favor of the alternative hypothesis. We now answer the question, how do we determine a good rejection region?

2.4 Methods of Evaluating Tests (C&B 8.3)

2.4.1 Types of Errors

When we select a rejection region for a hypothesis test, we need to examine the probabilities of making mistakes.

For the testing problem

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \in \Theta_0^C$$

there are two basic types of mistakes as displayed in the table below:

		Decision	
		Accept H_0	Reject H_0
Truth	H_0	Correct	*Type I Error*
	H_1	*Type II error*	Correct

A *type I error* is made if H_0 is rejected when H_0 is true. The *probability of a type I error* is denoted by α.

A *type II error* is made if H_0 is accepted when H_a is true. The *probability of a type II error* is denoted by β.

Traditionally, we are very conservative with our Type 1 error, think of "innocent until proven guilty" where we try hard not to make a Type 1 error by convicting an innocent person.

2.4.2 Power Function

Let $\mathcal{R}$ denote the rejection region for the test.

- If $\theta \in \Theta_0$ and $\mathbf{X} \in \mathcal{R}$ then the decision to reject is a mistake (type I error). The probability of a type I error depends on θ and is $P_\theta(\mathbf{X} \in \mathcal{R})$ for $\theta \in \Theta_0$.
- If $\theta \in \Theta_0^C$ and $\mathbf{X} \in \mathcal{R}^C$, then the decision to fail to reject is a mistake (type II error). The probability of a type II error depends on θ and is written $P_\theta(\mathbf{X} \in \mathcal{R}^C)$ for $\theta \in \Theta_0^C$.

Note for all θ, $P_\theta(\mathbf{X} \in \mathcal{R}^c) = 1 - P_\theta(\mathbf{X} \in \mathcal{R})$, so the function $P_\theta(\mathbf{X} \in \mathcal{R})$ contains all the information about the test with rejection region $\mathcal{R}$. It contains all the information about the error probabilities for a test. Notably:

$$P_\theta(\mathbf{X} \in \mathcal{R}) = \begin{cases} P_\theta(\text{ Type I Error }), & \theta \in \Theta_0 \\ 1 - P_\theta(\text{Type II Error }), & \theta \in \Theta_0^c. \end{cases}$$

Definition 8.3.1 This function of θ is called the *power function* of a test based on rejection $\mathcal{R}$:

$$\beta(\theta) = P_\theta(\mathbf{X} \in R)$$

Ideally we would use a test that has high power when $\theta \in \Theta_0^c$ and low power (type I error) when $\theta \in \Theta_0$. The perfect test would have a power function that is 0 for all $\theta \in \Theta_0$ and 1 for all $\theta \in \Theta_0^c$, but this can only be attained in trivial situations.

Example Suppose $X_1, \ldots, X_n \sim_{i.i.d} Bernoulli(p)$ and we want to test:

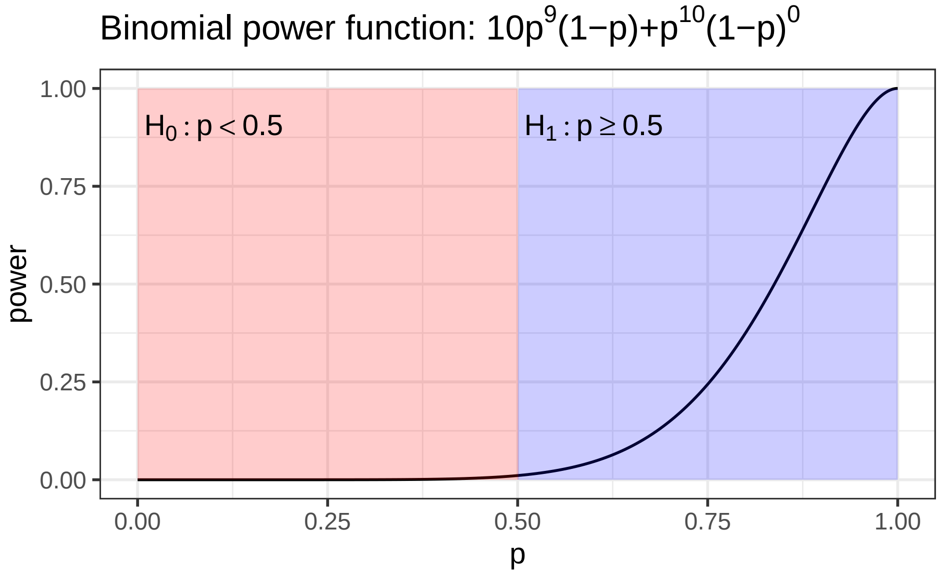
$$H_0 : p \leq 1/2 \text{ versus } H_1 : p > 1/2$$

Suppose we have $n = 10$ and form a rejection region, $\mathcal{R} = \{\mathbf{x} : \sum_{i=1}^{n} x_i \geq 9\}$. We can then calculate the power function:

$$P_p(\mathbf{X} \in \mathcal{R}) = P_p\left(\sum_{i=1}^{n} X_i \geq 9\right) = \binom{10}{9} p^9 (1-p)^1 + \binom{10}{10} p^{10} (1-p)^0.$$

A graph of this power function is given below. Note that it is an increasing function of p:

```
mydata <- tibble(p = seq(0,1, by=0.001)) %>% mutate(power = 10*p^9*(1-p)+1*p^10*(1-p)^0)
ggplot(mydata, aes(x=p, y=power))+geom_line()+
  ggtitle(TeX("Binomial power function: $10p^9(1-p)+p^{10}(1-p)^0$"))+
  geom_area(aes(y=ifelse(p > 0.5, 1, 0),x=p),fill="blue",alpha=0.2)+
  geom_area(aes(y=ifelse(p <= 0.5, 1, 0),x=p),fill="red",alpha=0.2)+
  annotate("text",x=0.1, y=.9,label="H[0]: p < 0.5", parse=TRUE)+
  annotate("text",x=0.6, y=.9,label="H[1]: p >= 0.5", parse=TRUE)
```



At $p = 1/2$, the power of the test is 0.011. For $p < 1/2$ the values are all less than 0.011. For $p > 1/2$ the values increase with p.

2.4.3 Size and level of tests

Generally we can't make both error probabilities as small as we'd like for a fixed sample size. A common approach is to put a bound on the type I error probability and find a good test (or 'best' test if you can) among those with P(type I error) $\leq$ bound.

Definition 8.3.5 For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a *size* α *test* if $\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$.
Definition 8.3.6 For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a *level* α *test* if $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$.

Different authors may use different definitions for size and level. In C&B, all size α tests are also level α tests. (For level α tests, the maximum type I error may be strictly less than α)

In planning a study (such as a clinical trial), researchers will often select a sample size, so that the test for the primary hypothesis has desired power for a specific value of $\theta \in \Theta_0^c$ and a specific size or level.

In the previous section we found the form of the rejection region of a variety of tests without specifying the details for the RR. That is, the test may have been of the form $\bar{\mathbf{X}} \geq c$ without specification of the value c. Using a condition for the size of the test, let us identify the value of c.

Example (Example 8.3.3 Normal power function) Let $X_1, \ldots, X_n \sim_{iid} Normal(\mu, \sigma^2), \sigma^2$ known. We choose our hypotheses:

$$H_0 : \mu \leq \mu_0 \text{ versus } H_1 : \mu > \mu_0$$

Let's use Theorem 8.2.4 which says we can use the distribution of our sufficient statistic $T(X) = \bar{\mathbf{X}}\ \mathcal{N}(\mu, \frac{\sigma^2}{n})$. So, the likelihood is

$$\mathcal{L}(\mu|\bar{x}) = (2\pi\sigma^2/n)^{-1} \exp\left(\frac{-(\bar{x}-\mu)^2}{2\sigma^2/n}\right)$$

If we maximize this likelihood in the full parameter space, we get the MLE $\widehat{\mu} = \bar{\mathbf{x}}$.

If $\mu_0 < \bar{x}$ we have $\widehat{\mu}_0 = \mu_0$. So the LRT in this case is

$$\lambda(\bar{x}) = \frac{(2\pi\sigma^2/n)^{-1} \exp\left(\frac{-(\bar{x}-\mu_0)^2}{2\sigma^2/n}\right)}{(2\pi\sigma^2/n)^{-1} \exp\left(\frac{-(\bar{x}-\bar{x})^2}{2\sigma^2/n}\right)} = \exp\left(\frac{-(\bar{x}-\mu_0)^2}{2\sigma^2/n}\right)$$

.

If $\mu_0 \geq \bar{x}$ we have $\widehat{\mu}_0 = \bar{x}$ and so the LRT $= 1$.

Hence, the LRT is

$$\lambda(\bar{x}) = \begin{cases} 1, & \mu_0 \geq \bar{x} \\ \exp\left(\frac{-(\bar{x}-\mu_0)^2}{2\sigma^2/n}\right), & mu_0 < \bar{\mathbf{x}} \end{cases}$$

and the rejection region is

$$
\begin{aligned}
\mathcal{R} &= \left\{ \mathbf{x} : \exp\left(\frac{-(\bar{x}-\mu_0)^2}{2\sigma^2/n} \right) \le c \right\} \\
&= \left\{ \mathbf{x} : |\bar{x}-\mu_0| \ge \sqrt{-2\sigma^2 \log(c)/n} \right\} \\
&= \left\{ \mathbf{x} : \bar{x} \ge \mu_0 + \sqrt{-2\sigma^2 \log(c)/n} \right\} \\
&= \{ \mathbf{x} : \bar{x} \ge k \}
\end{aligned}
$$

The LRT rejects *iff* $\bar{x} \ge k$. To obtain a size α test, choose k, so that $\sup_{\mu \le \mu_0} P(\bar{\mathbf{X}} \ge k) = \alpha$. Then the power function is:

$$
\begin{aligned}
\beta(\mu) = P_\mu(\bar{\mathbf{X}} \ge k) &= P_\mu \left(\frac{\bar{\mathbf{X}}-\mu}{\sigma/\sqrt{n}} \ge \frac{k-\mu}{\sigma/\sqrt{n}} \right) \\
&= P_\mu \left(Z \ge \frac{k-\mu}{\sigma/\sqrt{n}} \right) \\
&= 1 - \Phi \left(\frac{k-\mu}{\sigma/\sqrt{n}} \right)
\end{aligned}
$$

where Z is a standard normal variable and $\Phi(z)$ is the cdf of a Normal(0,1) variable. This is an increasing function of μ for any fixed k, n, and σ. So the supremum of this function with respect to μ under H_0 occurs at $\mu = \mu_0$. So, to determine the size we have:

$$
\alpha = \sup_{\mu \le \mu_0} \beta(\mu) = \beta(\mu_0) = 1 - \Phi \left(\frac{k-\mu_0}{\sigma/\sqrt{n}} \right)
$$

Now if we use the standard normal distribution to calculate z_α such that $1-\Phi(z_\alpha) = P(Z > z_\alpha) = \alpha$ we have:

$$
k = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}.
$$

For example, if $\alpha = 0.05$, we have $z_\alpha = 1.64$ (approximately) as $1-\Phi(1.64) = 1-P(Z \le 1.64) \approx 0.05$.

Now we have found an appropriate k to keep the size at α. So, our LRT rejection region is:

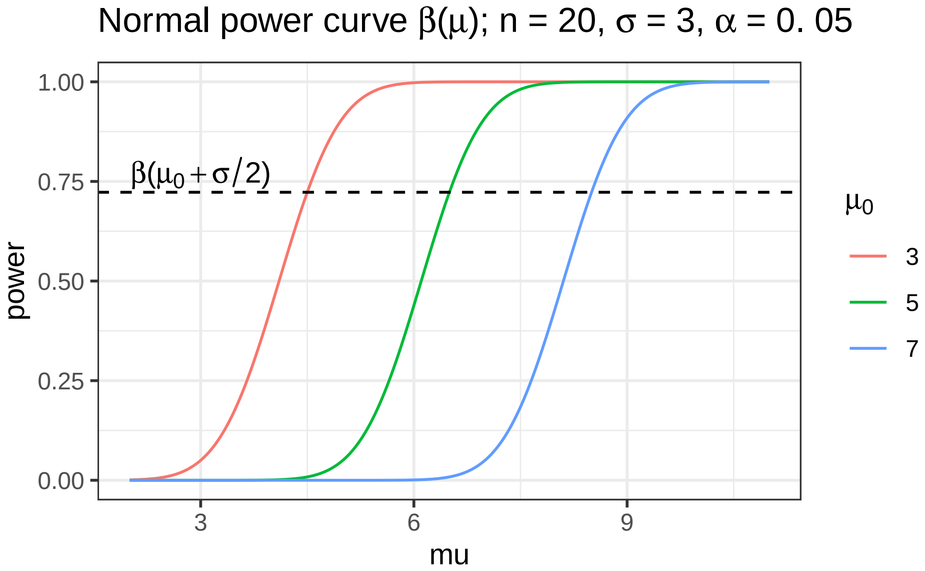
$$
\mathcal{R} = \{ \mathbf{x} : \bar{\mathbf{x}} \ge \mu_0 + z_\alpha \sigma/\sqrt{n} \}
$$

and our power function is:

$$
\beta(\mu) = P_\mu \left(\bar{\mathbf{X}} \ge \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} \right) = 1 - \Phi \left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha \right)
$$

We can plot $\beta(\mu)$ for various μ_0:

```r
ss <- 3 # sigma
nn <- 20
alpha <- 0.05
mydata <- crossing(mu0 = c(3,5,7), mu = seq(2,11,by=0.01)) %>%
  mutate(power = 1-pnorm((mu0-mu)/(ss/sqrt(nn))+qnorm(1-alpha)),
         mu0 = factor(mu0))
ggplot(mydata,aes(x = mu, y = power, color=mu0))+
  geom_line()+
  ggtitle(TeX("Normal power curve $\\beta(\\mu)$; n = 20, $\\sigma$ = 3, $\\alpha$ = 0.05"))+
  geom_hline(aes(yintercept=1-pnorm(qnorm(1-alpha)-sqrt(nn)/2)),color="black",lty=2)+
  annotate("text",x=3,y=.77,label=TeX("$\\beta(\\mu_0+\\sigma/2)$"))+
  scale_color_discrete(name=TeX("$\\mu_0$"))
```



How do we determine power under the alternative? When the size and μ_0 is fixed, power depends on μ and n. Suppose we care about a specific μ_1 in the alternative space, $\mu_1 = \mu_0 + \sigma/2$. Now we can calculate the power of our test under the alternative at this particular μ_1:

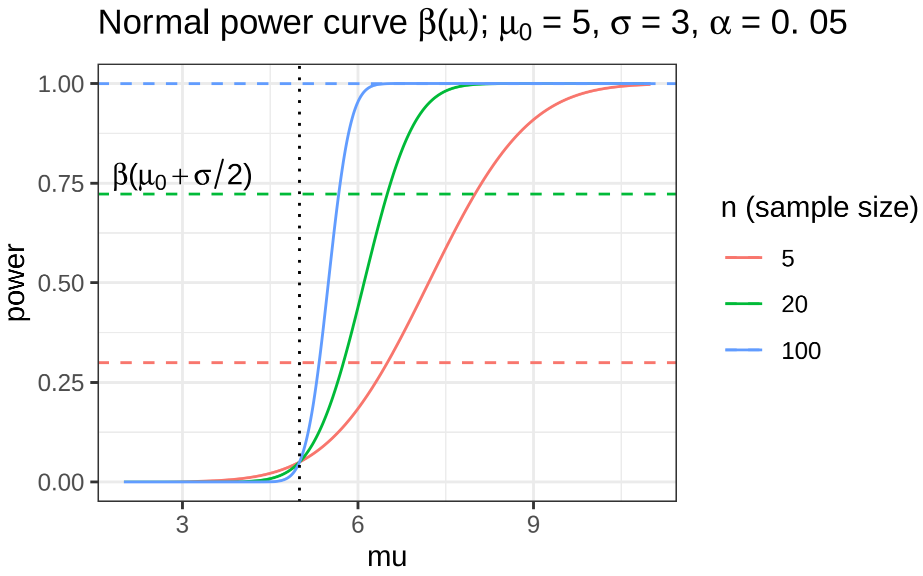
$$\beta(\mu_0 + \sigma/2) = 1 - \Phi\left(\frac{\mu_0 - (\mu_0 + \sigma/2))}{\sigma/\sqrt{n}} + z_\alpha\right)$$

$$= 1 - \Phi\left(z_\alpha - \sqrt{n}/2\right)$$

We see that this is now a function of n. So, for example, if we set our size to be $\alpha = 0.05$ then we have $z_\alpha = 1.64$, and if we have $n = 20$, then the power at $\mu_1 = \mu_0 + \sigma/2$ is $1 - \Phi(1.64 - \sqrt{20}/2) \approx 0.72$ which we can see in the plot above at $\mu_1 = \mu_0 + 3/2$ for each μ_0.

We can also plot power for a fixed μ_0 and various n:

```
ss <- 3 # sigma
alpha <- 0.05
mydata <- crossing(nn = c(5,20,100), mu0 = 5, mu = seq(2,11,by=0.01)) %>%
  mutate(power = 1-pnorm((mu0-mu)/(ss/sqrt(nn))+qnorm(1-alpha)),
         nn_fac = factor(nn))
ggplot(mydata,aes(x = mu, y = power, color=nn_fac))+
  geom_line()+
  ggtitle(
    TeX("Normal power curve $\\beta(\\mu)$; $\\mu_0$ = 5, $\\sigma$ = 3, $\\alpha$ = 0.05"))+
  geom_hline(aes(yintercept=1-pnorm(qnorm(1-alpha)-sqrt(nn)/2), color=nn_fac),lty=2)+
  geom_vline(xintercept = 5, lty=3)+
  annotate("text",x=3,y=.77,label=TeX("$\\beta(\\mu_0+\\sigma/2)$"))+
  scale_color_discrete(name="n (sample size)")
```



Now we see that $\beta(\mu_0 + \sigma/2)$ increases with n.

2.4.4 Unbiased tests

We would like a test we use to have at least as good power when the alternative is true as when the null hypothesis is true. In other words we want our power *under the alternative* to be higher than the *level* of our test (type I error). Tests that satisfy this property are referred to as unbiased.

Definition 8.3.9 A test with power function $\beta(\theta)$ is *unbiased* if $\beta(\theta') \geq \beta(\theta'')$ for every $\theta' \in \Theta_0^C$ and $\theta'' \in \Theta_0$.

Example (Example 8.3.10 Conclusion of 8.3.3 Normal power function) We saw earlier that the power function for the size α LRT of $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$ was monotone increasing in μ, so the power for any μ such that $\mu > \mu_0$ is greater than the power for any such μ

such that $\mu \leq \mu_0$. Hence the LRT test for this setting is unbiased. We see this in the plot of the power function above where the power function is always larger than $\alpha = 0.05$ when $\mu > \mu_0$.

2.4.5 Most powerful tests

There can be many size α tests for a given problem, also many unbiased tests for a given problem, so we will look for additional criteria to select a good test. The first consideration will be to find the most powerful level α test.

Definition 8.3.11 Let $\mathbb{C}$ be a class of tests for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^C$. A test in class $\mathbb{C}$, with power function $\beta(\theta)$, is a **uniformly most powerful (UMP)** class $\mathbb{C}$ test if $\beta(\theta) \geq \beta'(\theta)$ for every $\theta \in \Theta_0^C$ and every $\beta'(\theta)$ that is a power function of a test in class $\mathbb{C}$.

First we will consider the class $\mathbb{C}$ to be the class of all level α tests. We find the UMP level α test (when it exists) using the following theorem.

Theorem 8.3.12: Neyman-Pearson Lemma Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, where the pdf or pmf corresponding to θ_i is $f(\mathbf{x}|\theta_i), i = 0, 1$ using a test with *rejection region* $\mathcal{R}$ that satisfies (8.3.1):

$$\mathbf{x} \in \mathcal{R} \text{ if } f(\mathbf{x}|\theta_1) > kf(\mathbf{x}|\theta_0), \text{ and}$$

$$\mathbf{x} \in \mathcal{R}^c \text{ if } f(\mathbf{x}|\theta_1) < kf(\mathbf{x}|\theta_0),$$

for some $k \geq 0$, and (8.3.2):

$$\alpha = P_{\theta_0}(\mathbf{X} \in \mathcal{R}).$$

Then:

a) (*Sufficiency*) Any test that satisfies the above conditions (8.3.1 and 8.3.2) is a UMP level α test.
b) (*Necessity*) If there exists a test satisfying the above conditions (8.3.1 and 8.3.2) with $k > 0$, then every UMP level α test is a size α test and every UMP level α test satisfies 8.3.1 (except perhaps on a set that has probability zero when $\theta = \theta_0$ and $\theta = \theta_1$).

Proof: See C&B. Note that $\alpha = P_{\theta_0}(\mathbf{X} \in \mathcal{R})$ implies a size α test and hence a level α test since Θ_0 has only one point.

Note we are not saying anything about the form of the test when $f(\mathbf{x}|\theta_1) = kf(\mathbf{x}|\theta_0)$.

Note, also, we can write the rejection region as

$$\mathcal{R} = \left\{\mathbf{x} : \frac{f(\mathbf{x}|\theta_1)}{f(\mathbf{x}|\theta_0)} > k\right\}$$

and

$$\mathcal{R}^c = \left\{\mathbf{x} : \frac{f(\mathbf{x}|\theta_1)}{f(\mathbf{x}|\theta_0)} < k\right\}$$

provided we are not dividing by zero. This is an easier form of the lemma to work with.

Note in the Neyman-Pearson Lemma there is no requirement that the observations are *iid*. The pdf (pmf) could be from non-*iid* observations such as in a regression setting. We will see examples of this later.

As with the LRT test, we can also first reduce by sufficiency and then apply NP using the pdf (pmf) of the sufficient statistic.

Corollary 8.3.13: Consider the same hypothesis problem as in the Neyman-Pearson lemma. Suppose $T = T(\mathbf{X})$ is sufficient for θ and $g(t|\theta_i)$ is the pdf or pmf of T corresponding to $\theta_i, i = 0, 1$. Then, any test based on T with rejection region S is a UMP level α test if it satisfies

$$t \in S \text{ if } g(t|\theta_1) > kg(t|\theta_0), \text{ and}$$

$$t \in S^c \text{ if } g(t|\theta_1) < kg(t|\theta_0),$$

for some $k \geq 0$,

$$\alpha = P_{\theta_0}(T \in S). \text{ (8.3.5)}$$

Proof: By the factorization theorem, see C&B.

It can be easier to work with the pdf (or pmf) of a sufficient reduction of the data and hence this theorem is useful.

We can write the rejection region as

$$\mathcal{S} = \left\{\mathbf{x} : \frac{g(\mathbf{t}|\theta_1)}{g(\mathbf{t}|\theta_0)} > k\right\}$$

Example (Example 8.3.14 UMP binomial test: Let $X \sim Binomial(2, \theta)$. We want to test

$$H_0 : \theta = 1/2 \text{ versus } H_1 : \theta = 3/4.$$

We can use Neyman-Pearson lemma by calculating

$$\frac{f(\mathbf{x}|\theta_1)}{f(\mathbf{x}|\theta_0)}$$

for each possible value of $\mathbf{x}$, and finding a constant k that satisfies the lemma.

Calculating the ratios of the pmfs gives:

$$\frac{f(0|\theta = 3/4)}{f(0|\theta = 1/2)} = \frac{1}{4}, \frac{f(1|\theta = 3/4)}{f(1|\theta = 1/2)} = \frac{3}{4}, \text{ and } \frac{f(2|\theta = 3/4)}{f(2|\theta = 1/2)} = \frac{9}{4}$$

We need to choose a rejection region and k.

If we choose $3/4 < k < 9/4$, the NP lemma says that the rejection region $\mathcal{R} = \{X = 2\}$ is the UMP test with level $\alpha = \sup_{\theta=1/2} \beta(\theta) = P(X = 2|\theta = 1/2) = 1/4$. This is because for

$$\mathbf{x} \in \mathcal{R}, \frac{f(\mathbf{x}|\theta_1)}{f(\mathbf{x}|\theta_0)} > k$$

and for

$$\mathbf{x} \in \mathcal{R}^c,\ \frac{f(\mathbf{x}|\theta_1)}{f(\mathbf{x}|\theta_0)} < k.$$

On the other hand, if the rejection region is $\{X = 1 \text{ or } 2\}$, the NP lemma is satisfied with $1/4 < k < 3/4$ and so this test is UMP level $\alpha = P(X = 1 \text{ or } 2|\theta = 1/2) = 3/4$ test.

For extreme (trivial) cases, a $k < 1/4$ would correspond to the UMP test with rejection region $\{X >= 0\}$ and level $\alpha = P(X >= 0|\theta = 1/2) = 1$, and if $k > 9/4$ yields the UMP test with rejection region $X > 2\}$ and level $\alpha = P(X > 2|\theta = 1/2) = 0$.

Note: this shows that for a discrete distribution, the α level at which a test can be done is a function of the particular pmf of the data. No such problem arises in the continuous case. Any α level can be attained.

Example (Example 8.3.15 UMP normal test): Again we have $X_i \sim_{iid} N(\mu, \sigma^2)$ with σ^2 known. The same mean $\bar{\mathbf{X}}$ is the sufficient statistic for μ. Consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$ where $\mu_0 > \mu_1$. The inequality

$$g(\bar{\mathbf{x}}|\mu_1) > kg(\bar{\mathbf{x}}|\mu_0)$$

is equivalent to

$$\frac{\sqrt{n}}{\sigma\sqrt{2\pi}} e^{-n(\bar{\mathbf{x}}-\mu_1)^2/(2\sigma^2)} > k\frac{\sqrt{n}}{\sigma\sqrt{2\pi}} e^{-n(\bar{\mathbf{x}}-\mu_0)^2/(2\sigma^2)} \iff$$

$$(\bar{\mathbf{x}} - \mu_1)^2 < (2\sigma^2 \log(k))/n + (\bar{\mathbf{x}} - \mu_0)^2$$

$$\bar{\mathbf{x}} < \frac{(2\sigma^2 \log(k))/n - \mu_0^2 + \mu_1^2}{2(\mu_0 - \mu_1)}.$$

The right-hand side increases from $-\infty$ to ∞ as k increases from 0 to ∞. Thus, by Corollary 8.3.13, the test with rejection region $\bar{\mathbf{x}} < c$ is the UMP level α test, where $\alpha = P_{\mu_0}(\bar{\mathbf{X}} < c)$. If a particular α is specified, then the UMP test rejects H_0 if $\bar{\mathbf{X}} < c = \mu_0 - \sigma z_\alpha/\sqrt{n}$ (from above size calculations). This choice of c ensures that 8.3.5 is true.

2.4.6 Composite hypotheses

The NP lemma and corollary are concerned with *simple* hypotheses where only one distribution is specified by H_0 and H_1. In most realistic problems, the hypotheses of interest specify more than one possible distribution for the sample and are *composite* hypotheses.

For instance, suppose we are interested in a test of the form $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$, where, here we can see that each of the two hypotheses contain several parameter values (and hence several distributions). These types of hypotheses are composite hypotheses and in fact is a *one-sided* hypothesis.

The Neyman Pearson Lemma in many cases can be extended to composite hypotheses. In particular this may be the case for certain one-sided hypotheses.

For two-sided hypotheses, such as $H : \theta \neq \theta_0$, we cannot typically extend NP to get UMP tests.

For tests of one-sided hypotheses some distributions have a property that readily leads to UMP tests.

Definition 8.3.16 A family of pdfs or pmfs $\{f(x|\theta) : \theta \in \Theta\}$ for a univariate random variable X with real valued parameter θ has a **monotone likelihood ratio (MLR)** if for every $\theta_1 > \theta_2$

$$\frac{f(t|\theta_1)}{f(t|\theta_2)} = V(T(X), \theta_1, \theta_2)$$

where $V(T, \theta_1, \theta_2)$ is a monotone *non-decreasing* function of T.

Notes

- $c/0$ is defined as ∞ if $c > 0$
- C&B define MLR slightly differently, using the distribution of $T(X)$. This is equivalent due to the factorization and transformation theorems! C&B also define MLR as *either* monotone non-decreasing or non-increasing, but we need non-decreasing for the big theorem in this section.
- Many common families of distributions have an MLR. *Any regular exponential family with $f(x|\theta) = h(x)c(\theta)e^{w(\theta)T(x)}$ has an MLR if $w(\theta)$ is a non-decreasing function.* For example, normal (known variance, unknown mean), Poisson, and binomial all have an MLR.

Example: Bernoulli distribution Consider a family of Bernoulli distributions $\{B(p) : p \in [0, 1]\}$. The ratio of joint pmf of $X_1, \ldots, X_n$ is

$$\frac{f(\mathbf{X}|p_1)}{f(\mathbf{X}|p_2)} = \frac{p_1^{\sum X_i}(1-p_1)^{n-\sum X_i}}{p_2^{\sum X_i}(1-p_2)^{n-\sum X_i}} = \left(\frac{1-p_1}{1-p_2}\right)^n \left(\frac{p_1(1-p_2)}{p_2(1-p_1)}\right)^{\sum X_i}$$

For $p_1 > p_2$,

$$\frac{p_1(1-p_2)}{p_2(1-p_1)} > 1$$

and, therefore, the likelihood ratio is monotone increasing (non-decreasing) in $T(\mathbf{X}) = \sum_{i=1}^n X_i$.

Other ways to prove MLR: We could have also written this ratio as a function of T:

$$V(t) = \left(\frac{1-p_1}{1-p_2}\right)^n \left(\frac{p_1(1-p_2)}{p_2(1-p_1)}\right)^t$$

with derivative

$$\frac{\partial}{\partial t}V(t) = \left(\frac{1-p_1}{1-p_2}\right)^n t \left(\frac{p_1(1-p_2)}{p_2(1-p_1)}\right)^{t-1}$$

which is a product of three numbers ≥ 0, so the derivative is ≥ 0, and hence the ratio $V(t)$ is non-decreasing in t.

We could have also written the joint distribution as a regular exponential family where $w(p) = \log(p/(1-p))$, a non-decreasing function of p, since

$$f(\mathbf{x}|p) = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i} = (1-p)\left(\frac{p}{1-p}\right)^{\sum_i x_i} = (1-p)^n \exp\left[\left(\log\frac{p}{1-p}\right)\sum_i x_i\right].$$

Example: Exponential distribution Consider a family of Exponential distributions $\{Exp(\beta) : \beta \in (0,\infty)\}$. The ratio of joint pmf of $X_1,\ldots,X_n$ is

$$\frac{f(\mathbf{X}|\beta_1)}{f(\mathbf{X}|\beta_2)} = \frac{\beta_1^{-1}\exp(-\sum X_i/\beta_1)}{\beta_2^{-1}\exp(-\sum X_i/\beta_2)} = \left(\frac{\beta_2}{\beta_1}\right)\exp(\sum X_i/\beta_2 - \sum X_i/\beta_1) = \left(\frac{\beta_2}{\beta_1}\right)\exp\left[\frac{\beta_1-\beta_2}{\beta_1\beta_2}\sum X_i\right]$$

Since $\frac{\beta_1-\beta_2}{\beta_1\beta_2} > 0$, this ratio is monotone non-decreasing in $T(\mathbf{X}) = \sum X_i$.

Other ways to prove MLR: We could also have written this as a regular exponential family where $w(\beta) = -1/\beta$ which is a non-decreasing function of β. We could have also taken the derivative of the ratio with respect to t and shown it is ≥ 0.

Theorem 8.3.17 (Karlin-Rubin) Consider testing

$$H_0 : \theta \leq \theta_0 \text{ versus } H_1 : \theta > \theta_0.$$

Suppose, for any t_0,

1. T is a sufficient statistic for θ and
2. the family of pdfs or pmfs $\{f(x|\theta) : \theta \in \Theta\}$ of an MLR in T
3. $\alpha = P_{\theta_0}(T > t_0)$

Then the test with rejection region

$$\mathcal{R} = \{t : T > t_0\}$$

is a **UMP level α test**.

Conversely, if testing

$$H_0 : \theta \geq \theta_0 \text{ versus } H_1 : \theta < \theta_0,$$

with T a sufficient statistic for θ and $\{f(x|\theta) : \theta \in \Theta\}$ has an MLR in T. Then for any t_0, the $\mathcal{R} = \{t : T < t_0\}$ is a UMP level α test, where $\alpha = P_{\theta_0}(T < t_0)$.

Proof: In book. Essentially, the MLR allows us to use Neyman Pearson Lemma for any θ_1 in Θ_0 and θ_2 in Θ_1. We have the correct size test because the MLR gives us a power function $\beta(\theta) = P_\theta(T > t_0)$ that is also non-decreasing in θ.

Example: Exponential

Suppose we have $X_1, \dots X_n \sim \text{Exp}(\beta)$. We wish to test the hypothesis:

$$H_0 : \beta \leq \beta_0 \text{ versus } H_1 : \beta > \beta_0.$$

Can we find a UMP size α test?

From above we know that this family of distributions has an MLR in $T(\mathbf{X}) = \sum_{i=1}^n X_i$. So we can reject the null hypothesis with rejection region

$$\mathcal{R} = \left\{ \mathbf{x} : \sum_{i=1}^n X_i > t_0 \right\}$$

where t_0 is selected to satisfy the equation

$$\alpha = P_{\beta_0} \left(\sum_{i=1}^n X_i > t_0 \right)$$

To solve for t_0, we use the fact that $\sum_{i=1}^n X_i \sim \Gamma_{n,\beta} = \text{Gamma}(n, \beta)$ and so $t_0 = \gamma_{\alpha,n,\beta_0}$ which is the critical value of a $\text{Gamma}(n, \beta_0)$ distribution such that

$$\alpha = P_{\beta_0} \left(\Gamma_{n,\beta_0} > \gamma_{\alpha,n,\beta_0} \right)$$

Example: Normal UMP Test continued (Ex 8.3.18, continuation of Ex 8.3.15)

Again we have our normal data with variance known. Consider testing

$$H_0 : \mu \geq \mu_0 \text{ versus } H_1 : \mu < \mu_0$$

using the test with rejection region:

$$\mathcal{R} = \{\mathbf{x} : \bar{\mathbf{x}} < \mu_0 - \sigma z_\alpha / \sqrt{n}\}.$$

Let's **show this test is UMP and size** α.

As $\bar{\mathbf{X}}$ is sufficient and its distribution is $\text{Normal}(\mu, \sigma^2/n)$ with variance known and so is a one-dimensional exponential family with $w(\mu) = \mu/(2\sigma^2)$ increasing in μ, it follows from Karlin-Rubin Theorem (Thm 8.3.17) that the test is a UMP level α test in this problem.

We know the power of this test

$$\beta(\mu) = P_\mu(\bar{\mathbf{X}} < \mu_0 - \sigma z_\alpha / \sqrt{n}) = P\left(Z < \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha \right) = \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha \right)$$

is a decreasing function of μ since μ is a location parameter in the distribution of $\bar{\mathbf{X}}$ and so $\beta(\mu)$ is maximized in the null hypothesis space at μ_0 and $\beta(\mu_0) = \alpha$.

Example: Two-sided Normal Test (Example 8.3.19) Now consider the same Normality setting but with a two-sided test:

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu \neq \mu_0.$$

Can we find a UMP level α test in this setting?

Consider first the above one-sided test where we have $H_0^* : \mu \geq \mu_0$ versus $H_1^* : \mu < \mu_0$ and the UMP level α test rejects this H_0^* if $\bar{\mathbf{X}} < \mu_0 - z_\alpha \sigma/\sqrt{n}$. Let the power of this test be $\beta^*(\mu)$ and call this Test 1. Recall by NP Lemma that the best test must be of this form except on a set that has probability zero.

Conversely, if we wish to test $H_0^{**} : \mu \leq \mu_0$ against $H_1^{**} : \mu > \mu_0$ we can prove that the UMP level α test that rejects if $\bar{\mathbf{X}} > \mu_0 + z_\alpha \sigma/\sqrt{n}$. Let the power of this test be $\beta^{**}(\mu)$ and call it Test 2.

Now consider $\mu_1 > \mu_0$ which falls under the null hypothesis space of H_0^* and the alternative space H_1^{**}. Let's evaluate the power of the second one-sided test at μ_1:

$$\begin{aligned}
\beta^{**}(\mu_1) &= P_{\mu_1}(\bar{\mathbf{X}} > \mu_0 + z_\alpha \sigma/\sqrt{n}) \\
&= P_{\mu_1}\left(\frac{\bar{\mathbf{X}} - \mu_1}{\sigma/\sqrt{n}} > \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_\alpha\right) \\
&> P(Z > z_\alpha) \\
&= P(Z < -z_\alpha) \\
&> P_{\mu_1}\left(\frac{\bar{\mathbf{X}} - \mu_1}{\sigma/\sqrt{n}} < \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} - z_\alpha\right) \\
&= P_{\mu_1}(\bar{\mathbf{X}} < \mu_0 - z_\alpha \sigma/\sqrt{n}) \\
&= \beta^*(\mu_1)
\end{aligned}$$

So, the power of the Test 2 at μ_1 is greater than the power of Test 1 at μ_1. Hence, Test 1 cannot be the best test for all μ. But, recall the best test had to be of that form. So, no best test exists.

You can see in the plot below that although Test 1 and Test 2 have slightly higher powers than Test 3 for some parameter points, Test 3 has much higher power than Test 1 and Test 2 at other parameter points.

So, to get a "best" test for a setting like this we will have to **restrict the class of tests** we are willing to consider. The class we will consider here is the **class of unbiased tests**.

Note in this example for Test 1 the power is high when μ is small, but not when μ is large. In fact the power is $< \alpha$ when $\mu > \mu_0$. The opposite is true for Test 2. We'd like the power when the alternative hypothesis is true to be large and specifically greater than the power under the null hypothesis. Recall, we define the class of unbiased tests as tests with power function, $\beta(\theta)$, satisfying

$$\beta(\theta') \geq \beta(\theta''), \text{ for every, } \theta' \in \Theta_0^C, \theta'' \in \Theta_0.$$

Then we can often find a UMP unbiased level α test for one parameter problems involving composite hypotheses and for many two-sided problems as well.

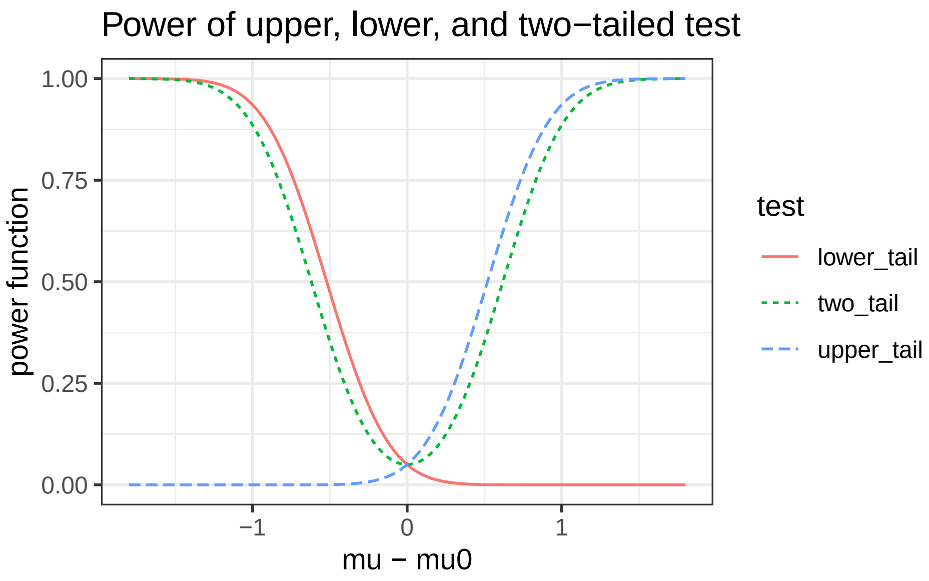
Example: Two-sided Normal Test Unbiased (Example 8.3.20)
When no UMP level α test exists within the class of all tests, we might try to find a UMP level α test within the class of unbiased tests.

The test which rejects $H_0 : \theta = \theta_0$ in favor of $H_1 : \theta \neq \theta_0$, if and only if

$$\bar{\mathbf{X}} > \mu_0 + z_{\alpha/2}\sigma/\sqrt{n}, \text{ or } \bar{\mathbf{X}} < \mu_0 - z_{\alpha/2}\sigma/\sqrt{n}$$

is a UMP unbiased level α test; that is, it is UMP in the class of unbiased tests.

```
ss <- 1 # sigma
alpha <- 0.05
nn <- 10
mydata <- tibble(mu = seq(-1.8,1.8,by=0.01)) %>%
  mutate(upper_tail = 1-pnorm((-mu)/(ss/sqrt(nn))+qnorm(1-alpha)),
         lower_tail = pnorm((-mu)/(ss/sqrt(nn))-qnorm(1-alpha)),
         two_tail = 1+pnorm((-mu)/(ss/sqrt(nn))-qnorm(1-alpha/2)) -
           pnorm((-mu)/(ss/sqrt(nn))+qnorm(1-alpha/2))
         ) %>%
  gather(key="test",value="power",-mu)
ggplot(mydata,aes(x = mu, y = power, color=test, lty = test))+
  geom_line()+
  xlab("mu - mu0")+
  ylab("power function")+
  ggtitle("Power of upper, lower, and two-tailed test")
```



2.4.7 p-values

As you know, it is more useful to report p-values than just what the size of the test is and what the test decision is (reject or not).

Definition 8.3.26 A *p-value* $p(\mathbf{X})$ is a **test statistic** satisfying $0 \leq p(\mathbf{x}) \leq 1$ for every sample point $\mathbf{x}$. Small values of $p(\mathbf{X})$ give evidence that H_1 is true.

A p-value is *valid* if, for every $\theta \in \Theta_0$ and every $0 \leq \alpha \leq 1$,

$$P_\theta(p(\mathbf{X}) \leq \alpha) \leq \alpha.$$

- if $p(\mathbf{X})$ is valid → can construct a level α test based on $p(\mathbf{X})$: the test that rejects H_0 if and only if $p(\mathbf{X}) \leq \alpha$ is a level α test
- reporting a test result via a p-value allows reader to choose the α he/she considers appropriate and compare reported $p(\mathbf{x})$ to α
- smaller p-value → stronger evidence for rejecting H_0
- p-value reports the results of a test on a more continuous scale, rather than just dichotomous decision "Accept H_0" or "Reject H_0."

Theorem 8.3.27 Let $W(\mathbf{X})$ be a test statistic such that large values of W give evidence that H_1 is true. For each sample point $\mathbf{x}$, define

$$p(\mathbf{x}) = \sup_{\theta \in \Theta_0} P_\theta(W(\mathbf{X}) \geq W(\mathbf{x})). \tag{8.3.9}$$

Then, $p(\mathbf{X})$ is a valid p-value.

Proof Fix $\theta \in \Theta_0$. Let $F_\theta(w)$ denote the cdf of $-W(\mathbf{X})$. Define

$$p_\theta(\mathbf{x}) = P_\theta(W(\mathbf{X}) \geq W(\mathbf{x})) = P_\theta(-W(\mathbf{X}) \leq -W(\mathbf{x})) = F_\theta(-W(\mathbf{x})).$$

- Hence, $p_\theta(\mathbf{X})$ is a random variable and is equal to $F_\theta(-W(\mathbf{X}))$.
- By the Probability Integral Transformation, $p_\theta(\mathbf{X}) \sim$Uniform(0,1) (or is stochastically equal to or greater than)
- Hence, $P_\theta(p_\theta(\mathbf{X}) \leq \alpha) \leq \alpha$ for every $0 \leq \alpha \leq 1$.

Now, since $p(\mathbf{x}) = \sup_{\theta' \in \Theta_0} p_{\theta'}(\mathbf{x}) \geq p_\theta(\mathbf{x})$ for every $\mathbf{x}$:

$$P_\theta(p(\mathbf{X}) \leq \alpha) \leq P_\theta(p_\theta(\mathbf{X}) \leq \alpha) \leq \alpha.$$

This is true for every $\theta \in \Theta_0$ and for every $0 \leq \alpha \leq 1 \Rightarrow p(\mathbf{X})$ is a valid p-value.

Example: One-sided normal p-value (Example 8.3.29) Assume we have $X_1, \ldots X_n \sim N(\mu, \sigma^2)$ with σ^2 unknown. Consider testing $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$. The LRT rejects H_0 for large values of $W(\mathbf{X}) = (\bar{\mathbf{X}} - \mu_0)/(S/\sqrt{n})$ (see C&B exercise 8.37) which has a Student's t distribution with $n-1$ degrees of freedom.

We can show that the supremum in the previous theorem always occurs at a parameter (μ_0, σ) and the value of σ does not matter:

For $\mu \leq \mu_0$ and any σ:

$$\begin{aligned} P_{\mu,\sigma}(W(\mathbf{X}) \geq W(\mathbf{x})) &= P_{\mu,\sigma}\left(\frac{\bar{\mathbf{X}} - \mu_0}{S/\sqrt{n}} \geq W(\mathbf{x})\right) \\ &= P_{\mu,\sigma}\left(\frac{\bar{\mathbf{X}} - \mu}{S/\sqrt{n}} \geq W(\mathbf{x}) + \frac{\mu_0 - \mu}{S/\sqrt{n}}\right) \\ &= P_{\mu,\sigma}\left(T_{n-1} \geq W(\mathbf{x}) + \frac{\mu_0 - \mu}{S/\sqrt{n}}\right) \\ &\leq P(T_{n-1} \geq W(\mathbf{x})). \end{aligned}$$

- The last inequality is true since $\mu_0 \geq \mu$ and $(\mu_0 - \mu)/(S/\sqrt{n})$ is a non-negative random variable.
- The probability does not depend on (μ, σ) so we can drop the subscript.
- $P(T_{n-1} \geq W(\mathbf{x})) = P_{\mu_0,\sigma}\left(\frac{\bar{\mathbf{X}} - \mu_0}{S/\sqrt{n}} \geq W(\mathbf{x})\right) = P_{\mu_0,\sigma}(W(\mathbf{X}) \geq W(\mathbf{x}))$, and since $(\mu_0, \sigma) \in \Theta_0$ this probability is included in the supremum in (8.3.9).

Thus, the p-value from (8.3.9) for this one sided t test is

$$p(\mathbf{x}) = P(T_{n-1} \geq W(\mathbf{x})) = P(T_{n-1} \geq (\bar{\mathbf{x}} - \mu_0)/(s/\sqrt{n})).$$

Note: Another method for defining a valid p-value, an alternative to using (8.3.9), involves conditioning on a sufficient statistic. If $S(\mathbf{X})$ is a sufficient statistic for the model $\{f(\mathbf{x}|\theta) : \theta \in \Theta_0\}$ and the null hypothesis is true, the conditional distribution of $\mathbf{X}|S = s$ does not depend on θ. So we define

$$p(\mathbf{x}) = P(W(\mathbf{X}) \geq W(\mathbf{x})|S = S(\mathbf{x})).$$

Similar to the proof in Theorem 8.3.27, but considering only the single distribution that is the conditional distribution $\mathbf{X}|S = s$, we see that, for any $0 \leq \alpha \leq 1$,

$$P(p(\mathbf{X}) \leq \alpha | S = s) \leq \alpha.$$

Thus, for any $\theta \in \Theta_0$, unconditionally we have

$$P_\theta(p(\mathbf{X}) \leq \alpha) = \sum_s P(p(\mathbf{X}) \leq \alpha | S = s) P_\theta(S = s) \leq \sum_s \alpha P_\theta(S = s) \leq \alpha.$$

Thus, $p(\mathbf{X})$ is a valid p-value. (Sums can be replaced by integrals for continuous S.)

- Fisher's Exact Test for comparing two proportions is a conditional test – using a conditional distribution to obtain the p-value.
- Conditional methods are often used in settings with "nuisance" parameters – parameters that are unknown but not of interest.

Example: Fisher's Exact Test (Example 8.3.30) Let $S_1 \sim$ *binomial*(n_1, p_1) and $S_2 \sim$ *binomial*(n_2, p_2) be independent. Consider testing

$$H_0 : p_1 = p_2 \text{ versus } H_1 : p_1 > p_2.$$

We can write the 2 × 2 table:

	Group 1	Group 2	Total
Condition 1	S_1	S_2	S
Condition 2	$n_1 - S_1$	$n_2 - S_2$	$n - S$
Total	n_1	n_2	n

Under H_0, if we let $p = p_1 = p_2$, the joint pmf of (S_1, S_2) is

$$\begin{aligned} f(s_1, s_2|p) &= \binom{n_1}{s_1} p^{s_1}(1-p)^{n_1-s_1} \binom{n_2}{s_2} p^{s_2}(1-p)^{n_2-s_2} \\ &= \binom{n_1}{s_1}\binom{n_2}{s_2} p^{s_1+s_2}(1-p)^{n_1+n_2-(s_1+s_2)} \end{aligned}$$

- Thus $S = S_1 + S_2$ is sufficient for p under H_0 (factorization).
- Now, p is unknown under H_0. So, we can't reduce this problem directly to one for which Neyman Pearson Lemma applies.
- However, since S is sufficient, we if we condition on S, we get a distribution that, under H_0, does not depend on p.
- Given the value of $S = s$, we can use S_1 as a test statistic and reject H_0 in favor of H_1 for large values of S_1 because large values of S_1 correspond to small values of $S_2 = s - S_1$.
- The conditional distribution, under H_0, of $S_1|S = s$ is hypergeometric$(n_1 + n_2, n_1, s)$. *(exercise 8.48)*

- We reject H_0 if S_1 is big and use the hypergeometric distribution to get the p-value:

$$P(S_1 \geq s_1 | S; n_1 + n_2, n_1).$$

  Here, the p-value is based on a conditional distribution.
- The conditional p-value is a sum of hypergeometric probabilities:

$$p(s_1, s_2) = P(S_1 \geq s_1 | S; n_1 + n_2, n_1) = \sum_{j=s_1}^{\min(n_1, s)} f(j|s)$$

- The test defined by this p-value is the Fisher's Exact Test, the best conditional test of the given size.

2.4.8 Bayesian Tests (C&B 8.2.2)

Recall that in the Bayesian paradigm, we place a prior distribution on the parameter θ and incorporate this with the likelihood given our sample data to obtain a posterior distribution. The classical/frequentist statistician considers θ to be fixed and consequently a hypothesis is either *true* or *false* and so the probabilities $P(H_0$ is true $|\mathbf{x})$ and $P(H_1$ is true $|\mathbf{x})$ are either 1 or 0 depending on the fixed value θ. However, in a Bayesian hypothesis testing problem, we believe θ to have a distribution and so the probability $P(H_0$ is true $|\mathbf{x}) = P(\theta \in \Theta_0 | \mathbf{x})$ (and analogously for H_1) is meaningful and may be computed.

To test hypotheses, a Bayesian approach compares $P(\theta \in \Theta_0 | \mathbf{X})$ to $P(\theta \in \Theta_0^C | \mathbf{X})$. Various rules can be devised. For example, we may decide to *accept* H_0 if $P(\theta \in \Theta_0 | \mathbf{X}) \geq P(\theta \in \Theta_0^C | \mathbf{X})$ and *reject* H_0 otherwise. Or, we may consider other criterion, such as *accept* H_0 if $P(\theta \in \Theta_0 | \mathbf{X})$ is quite large (say > 0.99).

Posterior Probability: Recall, by Bayes' Rule:

$$\begin{aligned} f(\theta|\mathbf{x}) &= \frac{f(\mathbf{x}|\theta)f(\theta)}{f(\mathbf{x})} \\ &\propto f(\mathbf{x}|\theta)f(\theta) \end{aligned}$$

where f is are the pmf's or pdf's of the random variables, and so we can use this to obtain the posterior distribution of $\theta|\mathbf{x}$ from the distributions of $\mathbf{x}|\theta$ and θ.

Example (*Bernoulli(p), p $\sim$ Uniform*)

Let $X_1, \ldots, X_n \sim$ *i.i.d.* Bernoulli(p) and let the prior on p be Uniform such that $p \sim \mathcal{U}(0, 1)$. We set the null and alternative hypotheses:

$$H_0 : p \leq 1/2 \text{ versus } H_1 : p > 1/2.$$

Now our pmf is

$$f(\mathbf{x}|p) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \prod_{i=1}^n x_i$$

and we can calculate the posterior

$$f(p|\mathbf{x}) = Beta(\sum_{i=1}^{n} x_i + 1, n - \sum_{i=1}^{n} x_i + 1).$$

We decide we will reject H_0 if $P(p \leq 1/2|\mathbf{x}) \leq 0.5$.

We obtain our data based on $n = 20$ and find $\sum_{i=1}^{n} x_i = 11$. Thus we have

$$P(Beta(12, 10) \leq 1/2) = 0.33.$$

So we would reject H_0 and conclude that $p > 1/2$.

However, if we were to observe $\sum_{i=1}^{n} x_i = 9$, we'd have $P(Beta(10, 12) \leq 1/2) = 0.67$ and would not reject H_0.

2.4.9 Summary of C&B Hypothesis Test Section

- The test statistic is a function of the sample i.e., $W(\mathbf{X})$
- The power function of the test is $\beta(\theta) = P_\theta(\mathbf{X} \in \text{Rejection Region})$
- A test is unbiased if $\beta(\theta_1) \geq \beta(\theta_2)$ for all $\theta_1 \in \Theta_A$ and $\theta_2 \in \Theta_0$
- The size of a test is $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$
- A p-value $p(\mathbf{X})$ is a test statistic and we are interested in valid p-values where $P_\theta(p(\mathbf{X}) \leq \alpha) \leq \alpha)$ so we may reject H_0 with an α level test: $\mathcal{R} : \{\mathbf{x} : p(\mathbf{x}) \leq \alpha\}$
- The Likelihood Ratio Test (LRT) is one useful way to create a test statistic

$$\lambda(\mathbf{x}) = \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta|\mathbf{x})}{\sup_{\theta \in \Theta} \mathcal{L}(\theta|\mathbf{x})}.$$

  - Rejection Region of the form: $\{\mathbf{x} : \lambda(\mathbf{x}) \leq c\}$
  - MLEs $(\widehat{\theta}_{MLE})$ help us to estimate the LRT since $\widehat{\theta}$ is the value of θ that maximizes the likelihood, but we also need to maximize the restricted $(\theta \in \Theta_0)$ likelihood
- Neyman-Pearson Lemma: When testing a null hypothesis against a simple alternative, tests based on a ratio of distributions (essentially, a ratio of likelihoods) are the most powerful.
- Karlin-Rubin Theorem: When testing against a one-sided hypothesis, having a Monotone Likelihood Ratio and one sided rejection region based on a sufficient statistic T gives us UMP test.
- Think about the Normal (known variance) one sided hypothesis problem.
  - We have a one sided null hypothesis $H_0 : \mu < \mu_0$ and $H_1 : \mu \geq \mu_0$
  - LRT rejects H_0 iff $\bar{\mathbf{X}} - \mu_0 \geq c$ for some c
  - If we want a size α test in this form, we determine that we need to reject H_0 iff $\bar{\mathbf{X}} \geq \mu_0 + z_\alpha \sigma/\sqrt{n}$.
  - Karlin Rubin tells us this is the Uniformly Most Powerful level α test for this hypothesis in the class of α level tests.
  - So we reject H_0 *iff* $\bar{\mathbf{X}} \in [\mu_0 + z_\alpha \sigma/\sqrt{n}, \infty)$. In Chapter 9 of *C&B* we will learn more about interval estimates and how they correspond to rejection regions.
  - As n becomes large, the critical point (i.e. the boundary of the rejection region) moves closer to θ_0. This will become interesting when we talk about asymptotic theory.

3 Interval Estimation (Chapter 9)

3.1 Definitions and Intro (9.1)

- In contrast to obtaining a single point estimate of a parameter of interest, we now want to construct an interval estimate (or more generally, a set estimate) for a parameter.
- A 95% confidence interval for the mean of a normal distribution based on sample data is a familiar example.
- Let's define what we mean by an interval estimate:

Definition 9.1.1: An **interval estimate** of a real-valued parameter θ is a pair of functions, $L(x_1, \ldots, x_n)$ and $U(x_1, \ldots, x_n)$, of a sample of data that satisfies $L(\mathbf{x}) \leq U(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$ where $\mathcal{X}$ is the sample space for $\mathbf{X}$. The **random interval** $[L(\mathbf{X}), U(\mathbf{X})]$ is called an *interval estimator.*

- If $\mathbf{X} = \mathbf{x}$ is observed, the inference $L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})$ is made.
- We write $[L(\mathbf{X}), U(\mathbf{X})]$ for an interval estimator of θ based on the random sample $\mathbf{X} = (X_1, \ldots, X_n)$ and $[L(\mathbf{x}), U(\mathbf{x})]$ based on the realized value of the interval.
- If we take $L(\mathbf{x}) = -\infty$ then $[-\infty, U(X_1, \ldots, X_n)]$ is a one-sided interval—an upper bound.
- Interval estimates can be closed intervals such as $[L(X_1, \ldots, X_n), U(X_1, \ldots, X_n)]$ or they can be open $(L(X_1, \ldots, X_n), U(X_1, \ldots, X_n))$.

- An advantage of intervals over point estimates is that we can attach a level of confidence to our interval.

Example (9.1.2 Interval Estimator) If we have a sample $X_1, X_2, X_3, X_4 \sim_{iid} \mathcal{N}(\mu, 1)$ an interval estimator of μ may be $[\bar{X} - 1, \bar{X} + 1]$. This means we will assert that μ is this interval.

- **Question** Why would we switch from a point estimator to an interval estimate? We are making the inference less precise, so what do we gain?
- **Answer** By giving up some precision in our estimate, we have gained some confidence, or assurance, that our assertion is correct.

Example $X_1, \ldots, X_n \sim_{iid} \mathcal{N}(\mu, \sigma^2)$, σ^2 known. A 95% CI for μ is

$$\left[\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n}\right].$$

What do we mean by 95% CI? How do we quantify our uncertainty about μ being in this interval?

Definition 9.1.4: For an interval estimator $[L(\mathbf{X}), U(\mathbf{X})] = [L(X_1, \ldots, X_n), U(X_1, \ldots, X_n)]$ of a parameter θ, the *coverage probability* of $[L(\mathbf{X}), U(\mathbf{X})]$ is the probability that the random interval $[L(\mathbf{X}), U(\mathbf{X})]$ covers the true parameter, θ. In symbols, it is denoted by either

$$P_\theta\left(\theta \in [L(\mathbf{X}), U(\mathbf{X})]\right) \text{ or } P\left(\theta \in [L(\mathbf{X}), U(\mathbf{X})] | \theta\right).$$

- Of course, we would like the *coverage probability* to be high. Note, that this coverage probability is a function of θ. We want the coverage probability to be high for all θ so we let the *confidence coefficient* be the minimum (or most technically infimum) over θ of the coverage probabilities.

Definition 9.1.4: For an interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$ of a parameter θ, the **confidence coefficient** of $[L(\mathbf{X}), U(\mathbf{X})]$ is the infimum of the coverage probabilities,

$$\inf_{\theta} P(\theta \in [L(\mathbf{X}), U(\mathbf{X})]).$$

- The **interval is random** not the parameter!
- Hence, probability statements involve the **probability with regard to $\mathbf{X}$**, not θ.
  - i.e. $P_\theta(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$ or $P_\theta(L(\mathbf{X}) \leq \theta, U(\mathbf{X}) \geq \theta)$ are statements about random $\mathbf{X}$
- An aside: In Bayesian approaches, we consider probability distributions on parameters, but not in the frequentist approaches we are discussing now.
- In general the coverage probability depends on θ, but in many cases, as in the normal mean example, the coverage probability is constant as a function of θ. In this case we do not need to worry about the "minimum" coverage.
- An interval estimate with a measure of confidence (usually confidence coefficient) are sometimes known as **confidence intervals.** C&B uses this interchangeably with **interval estimator.**
- More generally, we can describe **confidence sets** that are not continuous intervals.

An aside: Transformation Theorem 2.1.8 (simplified version) Let X have pdf $f_X(x)$, let $Y = g(X)$, and suppose (this is a loose summary of the requirements, page 53 in C&B for details)

- g is piecewise monotonic on the sample space $\mathcal{X}$. Essentially:
  - g is a 1:1 function
  - g^{-1} is a 1:1 function
- g^{-1} has a continuous derivative

Then

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$$

Read Chapter 2 in C&B for more details and proofs for monotonic g.

Example: Scale Uniform Interval Estimator (9.1.6) Let $X_1, \ldots, X_n$ be a random sample from a $Uniform(0, \theta)$ distribution, $\theta > 0$, and let $Y = \max\{X_1, \ldots, X_n\} = X_{(n)}$. Recall that Y is a sufficient statistic for θ (minimal sufficient), so it is reasonable to consider estimators based on Y. Suppose we consider two types of interval estimators: $[aY, bY]$ or $[Y + c, Y + d]$. Recall (see Example 7.3.13) that the density function of Y is

$$f_Y(y) = \frac{n}{\theta^n} y^{n-1} I(0 \leq y \leq \theta).$$

For the first type of interval we have

$$\begin{aligned} P_\theta(\theta \in [aY, bY]) &= P_\theta(aY \leq \theta \leq bY) \\ &= P_\theta\left(\frac{1}{b} \leq \frac{Y}{\theta} \leq \frac{1}{a}\right) \\ &= P_\theta\left(\frac{1}{b} \leq T \leq \frac{1}{a}\right) \\ &= \int_{1/b}^{1/a} nt^{n-1} dt = \left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n. \end{aligned}$$

since the pdf of T is $f_T(t) = nt^{n-1}, 0 \leq t \leq 1$ from the Transformation Theorem with $g(y) = y/\theta$ and $g^{-1}(t) = \theta t$, and $\frac{d}{dt} g^{-1}(t) = \theta$. This coverage probability is independent of the value of θ, and thus $\left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n$ is the confidence coefficient of the interval.

For the other interval, for $\theta \geq d$ we have:

$$\begin{aligned} P_\theta(\theta \in [Y + c, Y + d]) &= P_\theta(Y + c \leq \theta \leq Y + d) \\ &= P_\theta\left(1 - \frac{d}{\theta} \leq T \leq 1 - \frac{c}{\theta}\right) \\ &= \int_{1-d/\theta}^{1-c/\theta} nt^{n-1} dt \\ &= \left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n \end{aligned}$$

In this case, the coverage probability depends on θ and tends to zero as $\theta \to \infty$:

$$\lim_{\theta \to \infty} \left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n = 0.$$

So the confidence coefficient of this interval estimator is 0.

3.2 Methods for Finding Inverval Estimators (9.2)

3.2.1 Inverting a Test Statistic

Hypothesis testing naturally corresponds to interval estimation. In general, every confidence set corresponds to a test and vice versa.

Example: Inverting a Normal Test (Example 9.2.1)

We return to the Normal example, with σ known. Let $X_1, \ldots, X_n \sim_{iid} \mathcal{N}(\mu, \sigma^2)$ and we are testing the two-sided hypothesis

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu \neq \mu_0.$$

For a fixed α level, the most powerful unbiased test has rejection region

$$\mathcal{R} = \{\mathbf{x} : |\bar{x} - \mu_0| > z_{\alpha/2}\sigma/\sqrt{n}\}.$$

Note that H_0 is accepted for sample points in the complement of this region, or equivalently:

$$\mathcal{R}^C = \mathcal{A}(\mu_0) = \left\{\mathbf{x} : \bar{x} - z_{\alpha/2}\sigma/\sqrt{n} \leq \mu_0 \leq \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}\right\}.$$

This test has size α so $\Rightarrow P\left(H_0 \text{ is rejected } | \mu = \mu_0\right) = \alpha$, or, $P(H_0 \text{ is accepted } | \mu = \mu_0) = 1 - \alpha$. Hence,

$$P(\mathbf{X} \in \mathcal{A}(\mu_0)) = 1 - \alpha,$$

$\Rightarrow$

$$P\left(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} \leq \mu_0 \leq \bar{X} + z_{\alpha/2}\sigma/\sqrt{n} \mid \mu = \mu_0\right) = 1 - \alpha$$

But this probability statement is true for every μ_0. Hence, the statement

$$P\left(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}\right) = 1 - \alpha$$

is true.

The interval $C(x_1, \ldots, x_n) = [\bar{x} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}]$, obtained by *inverting* the acceptance region of the level α test, is a $1 - \alpha$ confidence interval.

Important Tautology:

$$\boxed{\mathbf{x} = \{x_1, \ldots, x_n\} \in A(\mu_0) \Leftrightarrow \mu_0 \in C(x_1, \ldots, x_n)}$$

Example: Let's go back to the interval above in *Example 9.2.1.* $X_1, \ldots, X_n \sim_{iid} \mathcal{N}(\mu, \sigma^2), \sigma^2$ known. A 95% CI for μ is

$$\left[\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n}\right].$$

Why? Because

$$P_\mu(\bar{X} - 1.96\sigma/\sqrt{n} \leq \mu \leq \bar{X} + 1.96\sigma/\sqrt{n}) = P(-1.96 \leq Z \leq 1.96) = 0.95,$$

for all μ. Hence the confidence coefficient for this interval is 0.95.

- A good hypothesis testing procedure leads to a good confidence interval and vice versa.
- Both tests and intervals ask the same question, but from a slightly different perspective.
- Both look for consistency between sample statistics and population parameters.
- Hypothesis test: fixes the parameters and asks what sample values (acceptance region) are consistent with that fixed value.
- Confidence set: fixes sample value and asks what parameter values (confidence interval) make this sample value most plausible
- If a test has *level* α then the confidence interval obtained by inversion has confidence coefficient *at least* $1 - \alpha$.
- We refer to this as a $1 - \alpha$ confidence interval or more generally a $1 - \alpha$ confidence set.

Theorem 9.2.2 For each $\theta_0 \in \Theta$, let $A(\theta_0)$ be the acceptance region of a level α test of $H_0 : \theta = \theta_0$. For each $\mathbf{x} \in \mathcal{X}$, define a set $C(\mathbf{x})$ in the parameter space by

$$C(\mathbf{x}) = \{\theta_0 : \mathbf{x} \in A(\theta_0)\}. \tag{9.2.1}$$

Then the random set $C(\mathbf{X})$ is a $1-\alpha$ confidence set.

Conversely, let $C(\mathbf{X})$ be a $1-\alpha$ confidence set. For any $\theta_0 \in \Theta$, define

$$A(\theta_0) = \{\mathbf{x} : \theta_0 \in C(\mathbf{x})\}.$$

Then $A(\theta_0)$ is the acceptance region of a level α test of $H_0 : \theta = \theta_0$.

Proof: For the first part:

$$\begin{aligned}
A(\theta_0) \text{ is the acceptance region of a level } \alpha \text{ test } &\Rightarrow P_{\theta_0}(\mathbf{X} \notin A(\theta_0)) \leq \alpha \\
&\Rightarrow 1 - P_{\theta_0}(\mathbf{X} \in A(\theta_0)) \leq \alpha \\
&\Rightarrow P_{\theta_0}\left[\mathbf{X} \in A(\theta_0)\right] \geq 1-\alpha \\
&\Rightarrow P_{\theta}\left[\mathbf{X} \in A(\theta)\right] \geq 1-\alpha \;\; (\theta_0 \text{ is arbitrary}) \\
&\Rightarrow P_{\theta}\left[\theta \in C(\mathbf{X})\right] = P_{\theta}\left[\mathbf{X} \in A(\theta)\right] \geq 1-\alpha \;\text{ with (9.2.1)} \\
&\Rightarrow C(\mathbf{X}) \text{ is a } 1-\alpha \text{ confidence set.}
\end{aligned}$$

For the second part, we write the Type I Error probability for the test of $H_0 : \theta = \theta_0$ with acceptance region $A(\theta_0)$:

$$P_{\theta}\left[\mathbf{X} \notin A(\theta)\right] = P_{\theta}\left[\theta \notin C(\mathbf{X})\right] \leq \alpha \Rightarrow \text{ level } \alpha \text{ test}$$

- **Important conclusion**: Theorem 9.2.2 shows that we must invert a family of tests, one for each value of $\theta_0 \in \Theta$, to obtain one confidence set.
- The first part of Theorem 9.2.2 is useful since it is relatively easy to construct a level α acceptance region but this helps us to construct a $1-\alpha$ confidence set, which is usually more difficult.
- In the Theorem we only specify $H_0 : \theta = \theta_0$. Usually we have to take in to account our H_A to decide the form of $A(\theta_0)$ and this will determine the shape of $C(\mathbf{x})$.
- Properties of inverted test carry over to the confidence set
  - Unbiased tests, when inverted, produce unbiased confidence sets.
  - We can use sufficient statistics to find a good test, so we can use sufficient statistics for good confidence tests.

Example: Inverting an LRT (9.2.3) Suppose that we want a confidence interval for the mean, λ of an $\exp(\lambda)$ population. We can obtain such an interval by inverting a level α test of

$$H_0 : \lambda = \lambda_0 \text{ versus } H_1 : \lambda \neq \lambda_0.$$

If we take a random sample $X_1, \ldots, X_n$, the LRT statistic is given by:

$$
\begin{aligned}
LRT &= \frac{\mathcal{L}(\lambda_0)}{\mathcal{L}(\widehat{\lambda}_{MLE})} \\
&= \frac{\frac{1}{\lambda_0^n} e^{-\sum x_i/\lambda_0}}{\frac{1}{\widehat{\lambda}_{MLE}^n} e^{-\sum x_i/\widehat{\lambda}_{MLE}}} \\
&= \frac{\frac{1}{\lambda_0^n} e^{-\sum x_i/\lambda_0}}{\frac{1}{(\sum x_i/n)^n} e^{-n}} \\
&= \left(\frac{\sum x_i}{n\lambda_0}\right)^n e^n e^{-\sum x_i/\lambda_0}.
\end{aligned}
$$

For fixed λ_0, the acceptance region is given by:

$$A(\lambda_0) = \left\{\mathbf{x} : \left(\frac{\sum x_i}{\lambda_0}\right)^n e^{-\sum x_i/\lambda_0} \geq k^*\right\},$$

where k^* is a constant chosen to satisfy

$$P_{\lambda_0}\left[\mathbf{X} \in A(\lambda_0)\right] = 1 - \alpha.$$

See plots in your book for a visualization of this region (pg 423). This is an interval in the sample space (a function of $\sum x_i$).

Inverting this acceptance region gives the $1 - \alpha$ confidence set:

$$C(\mathbf{x}) = \left\{\lambda : \left(\frac{\sum x_i}{\lambda}\right)^n e^{-\sum x_i/\lambda} \geq k^*\right\}.$$

This is an interval in the parameter space (a function of λ).

$C(\mathbf{x})$ depends on $\mathbf{x}$ only through $\sum x_i$. So the confidence interval can be expressed in the form:

$$C\left(\sum x_i\right) = \left\{\lambda : L\left(\sum x_i\right) \leq \lambda \leq U\left(\sum x_i\right)\right\},$$

where L and U are functions determined by the constraints that $P_{\lambda_0}(\mathbf{X} \in A(\lambda_0)) = 1 - \alpha$ and the constraint:

$$\left(\frac{\sum x_i}{L(\sum x_i)}\right)^n e^{-\sum x_i/L(\sum x_i)} = \left(\frac{\sum x_i}{U(\sum x_i)}\right)^n e^{-\sum x_i/U(\sum x_i)}. \tag{9.2.4}$$

If we set

$$\frac{\sum x_i}{L(\sum x_i)} = a \text{ and } \frac{\sum x_i}{U(\sum x_i)} = b, \tag{9.2.5}$$

where $a > b$ are constants, then the constraint (9.2.4) becomes

$$a^n e^{-a} = b^n e^{-b}, \tag{9.2.6}$$

which yields easily to numerical solution. To work out some details, let $n = 2$ and note that $\sum X_i \sim$ gamma(2,λ) and $\sum X_i/\lambda \sim$ gamma(2,1). Hence, from (9.2.5), the confidence interval becomes

$$\left\{\lambda : \frac{1}{a}\sum x_i \le \lambda \le \frac{1}{b}\sum x_i\right\},$$

where a and b satisfy

$$P_\lambda\left(\frac{1}{a}\sum X_i \le \lambda \le \frac{1}{b}\sum X_i\right) = P\left(b \le \frac{\sum X_i}{\lambda} \le a\right) = 1 - \alpha$$

and, from (9.2.6),

$$a^2 e^{-a} = b^2 e^{-b}.$$

Then

$$P\left(b \le \frac{\sum X_i}{\lambda} \le a\right) = \int_b^a te^{-t}dt$$

$$= e^{-b}(b+1) - e^{-a}(a+1).$$

If we want a 90% confidence interval, for example, we must simultaneously satisfy the probability condition and the constraint. Numerically solving these equations we get $a \approx 5.480, b \approx 0.441$ with a confidence coefficient of 0.90006. Thus,

$$P_\lambda\left(\frac{1}{5.480}\sum X_i \le \lambda \le \frac{1}{0.441}\sum X_i\right) = 0.90006.$$

Quick summary:

- The region obtained by inverting the LRT of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ (Definition 8.2.1) is of the form

  $$\text{accept } H_0 \text{ if } \frac{\mathcal{L}(\theta_0|\mathbf{x})}{\mathcal{L}(\widehat{\theta}|\mathbf{x})} \ge k(\theta_0),$$

  with the resulting confidence region

  $$\{\theta : \mathcal{L}(\theta|\mathbf{x}) \ge k'(\mathbf{x}, \theta)\},$$

  for some function k' that gives $1 - \alpha$ confidence.
- In some cases (i.e. normal and gamma distribution) the function k' will not depend on $\theta \Rightarrow$ likelihood region is interpreted as those values of θ for which the likelihood is highest
- the test inversion method is completely general since we can invert any test and obtain a confidence set (not just LRTs)

Example: Normal one-sided confidence bound (9.2.4) Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ population, μ and σ unknown but we wish to make inference about μ through a $1 - \alpha$ upper confidence bound for μ, of the form $C(\mathbf{x}) = (-\infty, U(\mathbf{x})]$.

Using Theorem 9.2., we can invert one-sided tests of $H_0 : \mu = \mu_0$ versus $H_1 : \mu < \mu_0$. (Note the direction of H_1 as compared to the "direction" of $C(\mathbf{x})$)

The size α LRT of H_0 versus H_1 has the rejection region:

$$\mathcal{R}(\mu_0) = \left\{\mathbf{x} : \frac{\bar{x} - \mu_0}{s/\sqrt{n}} < -t_{n-1,\alpha}\right\}$$

(similar to Example 8.2.6 in C&B). Thus the acceptance region for this test is:

$$\mathcal{A}(\mu_0) = \mathcal{R}(\mu_0)^C = \{\mathbf{x} : \bar{x} \geq \mu_0 - t_{n-1,\alpha}s/\sqrt{n}\}$$

and $\mathbf{x} \in A(\mu_0) \Leftrightarrow \bar{x} + t_{n-1,\alpha}s/\sqrt{n} \geq \mu_0$. We define

$$C(\mathbf{x}) = \{\mu_0 : \mathbf{x} \in A(\mu_0)\} = \{\mu : \bar{x} + t_{n-1,\alpha}s/\sqrt{n} \geq \mu\}\,.$$

By Theorem 9.2.2, the random set $C(\mathbf{X}) = (-\infty, \bar{X} + t_{n-1,\alpha}S/\sqrt{n}]$ is a $1-\alpha$ confidence set for μ. It is in the right form for an upper confidence bound. Inverting the one-sided test gave a one-sided confidence interval.

Example: Binomial one-sided confidence bound (9.2.5) Now we have a sequence of Bernoulli trials and we want to put a $1-\alpha$ lower confidence bound on p, where $X_1, \ldots, X_n \sim$ Bernoulli(p). We want the intervals to be of the form $(L(x_1, \ldots, x_n), 1]$ where $P_p(p \in L(X_1, \ldots, X_n), 1]) \geq 1 - \alpha$.

Since we want a one-sided interval with a lower confidence bound, we consider inverting the acceptance regions from tests of

$$H_0 : p = p_0 \text{ versus } H_1 : p > p_0.$$

To simplify things, we know we can base our test on the sufficient statistic $T = \sum_{i=1}^n X_i \sim$ binomial(n, p).

The binomial distribution has monotone likelihood ratio (Exercise 8.25) $\Rightarrow$ by the Karlin-Rubin Theorem, the test that rejects H_0 if $T > k(p_0)$ is the UMP test of its size.

For each p_0, we want to choose the constant $k(p_0)$ so that we have a level α test. We cannot get the size of the test to be exactly α, except for certain values of p_0, because of the discreteness of T. But we choose $k(p_0)$ so that the size of the test is as close to α as possible, without being larger.

Define $k(p_0)$ as the integer between 0 and n that simultaneously satisfies equation set (9.2.8):

$$\sum_{y=0}^{k(p_0)} \binom{n}{y} p_0^y (1-p_0)^{n-y} \geq 1 - \alpha$$

$$\sum_{y=0}^{k(p_0)-1} \binom{n}{y} p_0^y (1-p_0)^{n-y} < 1 - \alpha$$

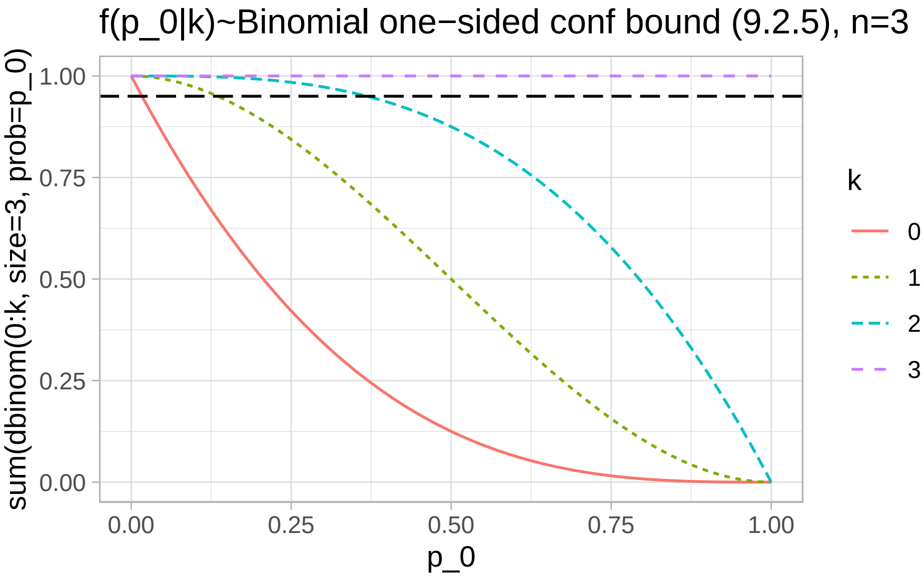
(this is as close to α as we can get)

Because of the MLR property of the binomial, for any fixed $k = 0, \ldots, n$, the quantity

$$f(p_0|k) = \sum_{y=0}^{k} \binom{n}{y} p_0^y (1-p_0)^{n-y}$$

is a decreasing function of p_0 (Exercise 8.26).

```r
library(tidyverse)
fdata <- crossing(k=0:3, p_0 = seq(0,1,by=.01)) %>%
  rowwise() %>%
  mutate(f_p_k = sum(dbinom(0:k, size=3, prob=p_0))) %>%
  ungroup() %>% mutate(k=factor(k))
ggplot(fdata, aes(x=p_0, y=f_p_k, color=k, lty=k))+
  geom_line()+
  geom_hline(yintercept=0.95, lty=5)+theme_light()+
  ggtitle("f(p_0|k)~Binomial one-sided conf bound (9.2.5), n=3")+
  ylab("sum(dbinom(0:k, size=3, prob=p_0)")
```



- Note that we choose $k(p_0)$ for each p_0, so $k(p_0)$ is an integer-valued step-function of p_0:
  - $f(0|0) = 1$, so $k(0) = 0$ and $f(p_0|0)$ remains above 1-α for an interval of values.
  - Then, at some point $f(p_0|0) = 1 - \alpha$ and for values of p_0 greater than this value, $f(p_0|0) < 1 - \alpha$. So, at this point, $k(p_0)$ increases to 1.
  - This pattern continues as p_0 increases and $k(p_0)$ increases one integer at a time: it is constant for a range of p_0, then it jumps to the next bigger integer.
- since $k(p_0)$ is a nondecreasing function of p_0, this gives the lower confidence bound
- solving the inequailities in (9.2.8) for $k(p_0)$ gives both the acceptance region of the test and the confidence set.
  - for each $p_0 : A(p_0) = \{t : t \leq k(p_0)\}$ where $k(p_0)$ satisfies (9.2.8)
  - for each value of t, the confidence set is $C(t) = \{p_0 : t \leq k(p_0)\}$.
- How do we define this interval explicitly in terms of p_0?
  - $k(p_0)$ is nondecreasing $\Rightarrow$ for a given observation $T = t, k(p_0) < t$ for all $p_0 < k^{-1}(t)$ where $k^{-1}(t)$ is some value we need to determine

- confidence set is $C(t) = \{p_0 : t \leq k(p_0)\} = \{p_0 : p_0 > k^{-1}(t)\}$ where

$$k^{-1}(t) = \sup \left\{ p : \sum_{y=0}^{t-1} \binom{n}{y} p^y (1-p)^{n-y} \geq 1 - \alpha \right\}$$

- The problem of binomial confidence bounds was first treated by Clopper and Pearson (1934) who obtained answers similar to this for the two-sided interval.
- Note this approach gives exact bounds. We instead could have used the normal approximation for large n to get an approximate 95% CI:

$$\widehat{p} \pm z_{\alpha/2} \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

3.2.2 Pivotal Quantities (9.2.2)

Remember that in Example (9.1.6) the coverage probability of $\{aY, bY\}$ did not depend on the value of θ while the $\{Y + c, Y + d\}$ interval did. This is because the coverage probability of $\{aY, bY\}$ could be expressed in terms of the quantity Y/θ, a random variable whose distribution does not depend on the parameter, a quantity known as a *pivotal quantity*, or *pivot*.

We want to find a random variable depending on both the data and the unknown parameter θ whose distribution does not depend on θ. Then, any probability statement about the pivotal quantity will not depend on θ.

Definition 9.2.6 A random variable $Q(\mathbf{X}, \theta) = Q(X_1, \ldots, X_n, \theta)$ is a *pivotal quantity* or *pivot* if the distribution of $Q(\mathbf{X}, \theta)$ is independent of all parameters. That is, if $\mathbf{X} \sim F(\mathbf{x}|\theta)$, then $Q(\mathbf{X}, \theta)$ has the same distribution for all values of θ.

- $Q(\mathbf{x}, \theta)$ will usually contain both parameters and statistics
- for any set $\mathcal{A}$, $P_\theta(Q(\mathbf{X}) \in \mathcal{A})$ cannot depend on θ

Example: Gamma Pivot (9.2.8) $X_1, \ldots, X_n \sim$ exponential(λ). Then $T = \sum X_i$ is a sufficient statistic for λ and $T \sim$ gamma(n, λ).

The gamma pdf is

$$\frac{1}{\Gamma(n)\lambda^n} t^{n-1} e^{-t/\lambda}$$

and note that t and λ appear together as t/λ and this gamma distribution is a scale family.

Thus, if $Q(T, \lambda) = 2T/\lambda$ and we let $Y = 2T/\lambda$ then by the transformation theorem,

$$g(t) = 2t/\lambda, \text{ and } g^{-1}(y) = \lambda y/2, \text{ and } \frac{d}{dy} g^{-1}(y) = \lambda/2$$

so

$$f_Y(y) = f_T(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| = \frac{1}{\Gamma(n)\lambda^n} \left(\frac{\lambda y}{2} \right)^{n-1} e^{-(\lambda y/2)/\lambda} |\lambda/2| = \frac{1}{\Gamma(n)2^n} e^{-y/2}$$

which does not depend on λ.

In fact, the quantity $Q(T,\lambda) = 2T/\lambda$ is a pivot with a gamma$(n,2) = \chi^2_{2n}$ distribution:

$$Q(T,\lambda) \sim \text{ gamma}(n, \lambda(2/\lambda)) = \text{gamma}(n,2) = \chi^2_{2n}$$

Note that $Q(T,\lambda) = T/\lambda$ is also a pivot with distribution $\Gamma(n,1)$.

How to find the pivot?

- There is no all purpose strategy for finding pivots.
- However, it is relatively easy to find pivots for location or scale parameters. In general, differences are pivotal for location parameters, while ratios (or products) are pivotal for scale problems.
- looking at the distribution can give us a clue about the pivot, for instance t/λ or $(\bar{x}-\mu)/\sigma$.
- In general, suppose the pdf of a statistic $T, f(t|\theta)$ can be expressed in the form:

  $$f(t|\theta) = g(Q(t,\theta)) \left| \frac{\partial}{\partial t} Q(t,\theta) \right|$$

  for some function g and some monotone function Q (monotone in t for each θ), then Theorem 2.1.5 can be used to show that $Q(T,\theta)$ is a pivot.

Using the pivot: If $Q(\mathbf{X},\theta)$ is a pivot, then for a specified value of α we can find numbers a and b, which do not depend on θ to satisfy:

$$P_\theta(a \le Q(\mathbf{X},\theta) \le b) \ge 1-\alpha.$$

Then, for each $\theta_0 \in \Theta$,

$$A(\theta_0) = \{\mathbf{x} : a \le Q(\mathbf{x},\theta_0) \le b)$$

is the acceptance region for a level α test of $H_0 : \theta = \theta_0$. Using Theorem 9.2.2, we invert these tests (for each θ_0) to obtain:

$$C(\mathbf{x}) = \{\theta_0 : a \le Q(\mathbf{x},\theta_0) \le b\},$$

and $C(\mathbf{X})$ is a 1-α confidence set for θ.

Example: Continuation of Gamma Example 9.2.8 We inverted a LRT to obtain the confidence interval for the mean λ of an exponential(λ) pdf for $H_0 : \lambda = \lambda_0$ versus $H_1 : \lambda \ne \lambda_0$.

Now we also see that if we have $X_1, \ldots, X_n$, we can define $T = \sum X_i$ and $Q(T,\lambda) = 2T/\lambda \sim \chi^2_{2n}$

If we choose constants a and b to satisfy $P(a \le \chi^2_{2n} \le b) = 1-\alpha$ then

$$P_\lambda\left(a \le \frac{2T}{\lambda} \le b\right) = P_\lambda(a \le Q(T,\lambda) \le b) = P(a \le \chi^2_{2n} \le b) = 1-\alpha.$$

Inverting the set $A(\lambda) = \{t : a \le 2t/\lambda \le b\} \Rightarrow C(t) = \{\lambda : 2t/b \le \lambda \le 2t/a\}$ which is a $1-\alpha$ confidence interval.

For example, if $n = 10$, then consulting a table of χ^2 cutoffs gives a 95% CI of $\{\lambda : 2T/34.17 \le \lambda \le 2T/9.59\}$.

Example: Normal pivotal interval (9.2.10)

This is a location problem. If we have $X_1, \ldots, X_n \sim_{iid} \mathcal{N}(\mu, \sigma^2)$, we know the distribution of $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ is $\mathcal{N}(0,1)$ so that statistic is a pivot.

if σ^2 is known:

$$P\left(-a \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq a\right) = P(-a \leq Z \leq a),$$

$$\Rightarrow C(\bar{x}) = \left\{\mu : \bar{x} - a\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + a\frac{\sigma}{\sqrt{n}}\right\}$$

If σ^2 is unknown we use the sample estimate of σ, S:

$$P\left(-a \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq a\right) = P(-a \leq T_{n-1} \leq a),$$

$$\Rightarrow C(\bar{x}) = \left\{\mu : \bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right\}$$

Now if we want an interval estimate of σ, we know that $(n-1)^2 S^2/\sigma^2 \sim \chi^2_{n-1}$ so $(n-1)^2 S^2/\sigma^2$ is also a pivot. Thus:

$$P\left(a \leq \frac{(n-1)S^2}{\sigma^2} \leq b\right) = P(a \leq \chi^2_{n-1} \leq b) = 1 - \alpha$$

$$\Rightarrow \left\{\sigma : \sqrt{\frac{(n-1)s^2}{b}} \leq \sigma \leq \sqrt{\frac{(n-1)s^2}{a}}\right\}.$$

We can choose $a = \chi^2_{n-1,1-\alpha/2}$ and $b = \chi^2_{n-1,\alpha/2}$ so that the probability is equally split between tails. However, the chi-square distribution is skewed so this is not actually the optimal choice (we will explore this later in 9.3).

We constructed confidence intervals for μ and σ separately, but if we want a simultaneous confidence interval we can use the Bonferroni Inequality (exercise 9.14).

3.2.3 Pivoting the CDF

- We saw that a pivot, Q, leads to a confidence set of the form:

  $$C(\mathbf{x}) = \{\theta_0 : a \leq Q(\mathbf{x}, \theta_0) \leq b\}.$$

  This will be an interval if $Q(\mathbf{x}, \theta)$ is a monotone function of θ for every $\mathbf{x}$.
- If possible, constructing a confidence set based on a LRT will give a good (maybe not optimal) set. If it is too difficult to invert, the method of pivoting the CDF can be applied and will usually produce a reasonable set.

How to pivot the CDF?

- Suppose we base inference on a statistic, T, (ideally sufficient) for θ, where $F_T(t|\theta)$ is the cdf for T. Suppose that $F_T(t|\theta)$ is monotone in θ.
- Note, if $F_T(t|\theta)$ is increasing (decreasing) in θ for every t in the sample space of T, we say that T is stochastically decreasing (increasing) in θ.
- Recall the Probability Integral Transformation, which tells us that the random variable $F_T(T|\theta)$ is Uniform(0,1), a pivot.
- Thus, if $\alpha_1 + \alpha_2 = \alpha$, an α-level acceptance region of the hypothesis $H_0 : \theta = \theta_0$ is

  $$\{t : \alpha_1 \leq F_T(t|\theta_0) \leq 1 - \alpha_2\},$$

  with associated confidence set

  $$\{\theta : \alpha_1 \leq F_T(t|\theta) \leq 1 - \alpha_2\}.$$

- To guarantee that the confidence set is an interval, we need to have $F_T(t|\theta)$ to be monotone in θ

Theorem 9.2.12 Pivoting a *continuous* cdf Let T be a statistic with continuous cdf $F_T(t|\theta)$. Let $\alpha_1 + \alpha_2 = \alpha$ with $0 < \alpha < 1$ be fixed values. Suppose that for each $t \in \mathcal{T}$, the functions $\theta_L(t)$ and $\theta_U(t)$ can be defined as follows

i. If $F_t(t|\theta)$ is a decreasing function of θ for each t, define $\theta_L(t)$ and $\theta_U(t)$ by

$$F_T(t|\theta_U(t)) = \alpha_1, \; F_T(t|\theta_L(t)) = 1 - \alpha_2.$$

ii. If $F_T(t|\theta)$ is an increasing function of θ for each t, define $\theta_L(t)$ and $\theta_U(t)$ by

$$F_T(t|\theta_U(t)) = 1 - \alpha_2, \; F_T(t|\theta_L(t)) = \alpha_1.$$

Then the random interval $[\theta_L(t), \theta_U(t)]$ is a $1 - \alpha$ confidence interval for θ.

Proof (of part (i)):

First recall the general idea behind the proof of the Probablity Integral Transform, where $Y = F_T(T)$:

$$F_Y(y) = P(Y \leq y) = P(F_T(T) \leq y) = P(T \leq F_T^{-1}(y)) = F_T(F_T^{-1}(y)) = y$$

So Y has cdf $F_Y(y) = y$, which is the cdf of Uniform(0,1).

Now, assume we have constructed the 1-α acceptance region:

$$\{t : \alpha_1 \leq F_T(t|\theta_0) \leq 1 - \alpha_2\}.$$

Since $F_T(T|\theta)$ is decreasing in θ for all $t \Rightarrow$

$$F_T(t|\theta) < \alpha_1 \Leftrightarrow \theta > \theta_U(t).$$

Similarly

$$F_T(t|\theta) > 1 - \alpha_2 \Leftrightarrow \theta < \theta_L(t)$$

Hence:

$$\begin{aligned}P(\theta_L(T) \leq \theta \leq \theta_U(T)) &= 1 - \{P(\theta > \theta_U(T)) + P(\theta < \theta_L(T))\} \\ &= 1 - \{P_\theta(F_T(T|\theta) < \alpha_1) + P_\theta(F_T(T|\theta) > 1 - \alpha_2)\} \\ &= 1 - (\alpha_1 + \alpha_2) = 1 - \alpha\end{aligned}$$

The proof of part (ii) is similar.

Notes:

- Note: we can assume $\alpha_1 = \alpha_2 = \alpha/2$ but this may not be optimal split.
- If we want a one-sided interval we can choose $\alpha_1 = 0$ or $\alpha_2 = 0$.

Example 9.2.13 Location exponential interval

Let $X_1, \ldots, X_n \sim_{iid} f(x|\mu) = e^{-(x-\mu)} I_{[\mu,\infty)}(x)$. Then $Y = min\{X_1, \ldots, X_n\} = X_{(1)}$ is sufficient for μ with pdf $f_Y(y|\mu) = ne^{-n(y-\mu)} I_{[\mu,\infty)}(y)$.

We have a decreasing function so we can fix α and define $\mu_L(y)$ and $\mu_U(y)$ to satisfy:

$$F_Y(y|\mu_U(y)) = \frac{\alpha}{2}, \text{ and } F_Y(y|\mu_L(y)) = 1 - \frac{\alpha}{2}$$

$$\int_{\mu_U(y)}^{y} ne^{-n(u-\mu_U(y))} du = \frac{\alpha}{2}, \text{ and } \int_{y}^{\infty} ne^{-n(u-\mu_L(y))} du = \frac{\alpha}{2}$$

Evaluate the integrals:

$$1 - e^{-n(y-\mu_U(y))} = \frac{\alpha}{2}, \text{ and } e^{-n(y-\mu_L(y))} = \frac{\alpha}{2}$$

then solve:

$$\mu_U(y) = y + \frac{1}{n}\log\left(1 - \frac{\alpha}{2}\right), \text{ and } \mu_L(y) = y + \frac{1}{n}\log\left(\frac{\alpha}{2}\right)$$

So this random interval is a 1-α confidence interval for μ:

$$C(Y) = \left\{\mu : Y + \frac{1}{n}\log\left(\frac{\alpha}{2}\right) \leq \mu \leq Y + \frac{1}{n}\log\left(1 - \frac{\alpha}{2}\right)\right\}$$

- Note that we only need to solve the CDF equations for the value of the statistics actually observed.
- We may solve them analytically or numerically.
- Now consider the discrete case:

Theorem 9.2.14 Pivoting a *discrete* cdf Let T be a discrete statistic with cdf $F_T(t|\theta) = P(T \leq t|\theta)$. Let $\alpha_1 + \alpha_2 = \alpha$ with $0 < \alpha < 1$ be fixed values. Suppose that for each $t \in \mathcal{T}$, the functions $\theta_L(t)$ and $\theta_U(t)$ can be defined as follows

i. If $F_t(t|\theta)$ is a decreasing function of θ for each t, define $\theta_L(t)$ and $\theta_U(t)$ by

$$P(T \leq t|\theta_U(t)) = \alpha_1, \; P(T \geq t|\theta_L(t)) = \alpha_2.$$

ii. If $F_T(t|\theta)$ is an increasing function of θ for each t, define $\theta_L(t)$ and $\theta_U(t)$ by

$$P(T \geq t|\theta_U(t)) = \alpha_1,\ P(T \leq t|\theta_L(t)) = \alpha_2.$$

Then the random interval $[\theta_L(t), \theta_U(t)]$ is a $1-\alpha$ confidence interval for θ.

Proof: See book.

Example 9.2.15 Poisson Interval Estimator Let $X_1, \ldots, X_n \sim_{iid}$ Poisson(λ) and define $T = \sum_{i=1}^n X_i$. Then T is sufficient for λ and $Y \sim$ Poisson($n\lambda$).

Let $\alpha_1 = \alpha_2 = \alpha/2$ in the above theorem. The cdf is a decreasing function of λ for each t since the cdf is

$$P(T \leq t|\lambda) = \sum_{k=0}^{t} e^{-n\lambda} \frac{(n\lambda)^k}{k!}$$

and has a negative derivative for all values of $\lambda > 0$:

$$\frac{d}{d\lambda} P(T \leq t|\lambda) = \sum_{k=0}^{t} -ne^{-n\lambda} \frac{kn^k \lambda^{(k-1)}}{k!} < 0$$

So, if $T = t_0$ is observed, we need to solve λ in the equations to obtain the lower and upper limits λ_L and λ_U:

$$P(T \leq t_0|\lambda_U) = \sum_{k=0}^{t_0} e^{-n\lambda_U} \frac{(n\lambda_U)^k}{k!} = \frac{\alpha}{2} \text{ and } P(T \geq t_0|\lambda_L) = \sum_{k=t_0}^{\infty} e^{-n\lambda_L} \frac{(n\lambda_L)^k}{k!} = \frac{\alpha}{2} \qquad (9.2.16)$$

How do we solve these equations? First, recall: If G is a gamma(a, b) random variable, where a is an integer, then for any g,

$$P(G \leq g) = P(H \geq a) \text{ where } H \sim Poisson(g/b).$$

Hence,

$$
\begin{aligned}
\alpha/2 &= \sum_{k=0}^{t_0} e^{-n\lambda_U} \frac{(nk)^k}{k!} \\
&= P(Y \leq t_0 | \lambda_U) \\
&= P(Pois(n\lambda_U) < t_0 + 1) \\
&= 1 - P(Pois(n\lambda_U) \geq t_0 + 1) \\
&= 1 - P(Gamma(t_0 + 1, 2) \leq 2n\lambda_U) \\
&= 1 - P(\chi^2_{2(t_0+1)} \leq 2n\lambda_U) \\
&= P(\chi^2_{2(t_0+1)} > 2n\lambda_U)
\end{aligned}
$$

Then solving for the critical value:

$$2n\lambda_U = \chi^2_{2(t_0+1),\alpha/2}$$

Solving for λ_U:

$$\lambda_U = \frac{1}{2n}\chi^2_{2(t_0+1),\alpha/2}$$

Applying the identity to the other equation in 9.2.16 yields:

$$\frac{\alpha}{2} = P(\chi^2_{2t_0} < 2n\lambda_L) \Rightarrow 1 - \frac{\alpha}{2} = P(\chi^2_{2t_0} \geq 2n\lambda_L).$$

Then solving for the critical value:

$$2n\lambda_L = \chi^2_{2t_0,1-\alpha/2}$$

and so

$$\lambda_L = \frac{1}{2n}\chi^2_{2t_0,1-\alpha/2}$$

and a $1 - \alpha$ confidence interval for λ is:

$$\left\{\lambda : \frac{1}{2n}\chi^2_{2t_0,1-\alpha/2} \leq \lambda \leq \frac{1}{2n}\chi^2_{2(t_0+1),\alpha/2}\right\}.$$

Suppose $t_0 = 4$ and $\alpha = 0.05$ and $n = 20$, then the confidence interval is:

```
nn = 20
t0 = 4
alpha = .05

lambda_l <- qchisq(p=1-alpha/2, df=2*t0, lower.tail = FALSE)/(2*nn)
lambda_u <- qchisq(p=alpha/2, df=2*(t0+1), lower.tail = FALSE)/(2*nn)
c(lambda_l, lambda_u)
```

```
#> [1] 0.05449327 0.51207943
```

The 95% CI for λ is: [0.054, 0.512]

For $n = 100$:

```
nn = 100
t0 = 4
alpha = .05

lambda_l <- qchisq(p=1-alpha/2, df=2*t0, lower.tail = FALSE)/(2*nn)
lambda_u <- qchisq(p=alpha/2, df=2*(t0+1), lower.tail = FALSE)/(2*nn)
c(lambda_l, lambda_u)
```

```
#> [1] 0.01089865 0.10241589
```

The 95% CI for λ is: [0.011, 0.102]

Similar derivations can be done for negative binomial and binomial distributions. The graph of coverage probabilities for binomial confidence intervals is given in Figure 9.2.5.

3.3 Methods of Evaluating Interval Estimators (9.3)

What makes a good interval? Small size and large coverage

- $(-\infty, \infty)$ has coverage probability 1
- in order to optimize, need to know how to measure these quantities
- sometimes a function of the parameter
- coverage usually measured by confidence coefficient (infimum of converge probabilities)
- size of an interval usually means length

Example 9.3.1 Optimimizing Length Let $X_1, \ldots, X_n \sim_{iid} \mathcal{N}(\mu, \sigma^2)$ where σ is known.

Using the pivotal quantity

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

which has a distribution that does not depend on μ, we know (section 9.2.2 Pivotal Quantities) that if we choose a and b such that: $P(a \leq Z \leq b) = 1 - \alpha$ we will have the $1 - \alpha$ confidence interval

$$C(\mu) = \left\{ \mu : \bar{x} - b\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} - a\frac{\sigma}{\sqrt{n}} \right\}.$$

This gives $1 - \alpha$ coverage, and a $1 - \alpha$ confidence coefficient.

How can we minimize the length?

$$length(C(\mu)) = \bar{x} - a\frac{\sigma}{\sqrt{n}} - \bar{x} + b\frac{\sigma}{\sqrt{n}} = \frac{(b-a)\sigma}{\sqrt{n}}$$

So we need to minimize $b - a$.

In Example 9.2.1 we chose $a = -z_{\alpha/2}$ and $b = z_{\alpha/2}$ but we did not consider optimality. We can take any other combination of values $a = z_a$ and $b = z_b$ and such that $P(a \leq Z \leq b) = z_b + 1 - z_a = 1 - \alpha$, and in fact if we let $\alpha = 0.1$ then we have many options for 90% intervals (see table in C&B, page 441).

Looking at those examples we see that splitting the probability α equally is an optimal strategy, as $a = -1.65 = -z_{.05}$ and $b = 1.65 = z_{.05}$ gives the shortest interval.

- splitting α equally does not always give the optimal length
- in the previous case, the height of the pdf is the same at $-z_{\alpha/2}$ and $z_{\alpha/2}$
- we can prove a theorem that demonstrates a general version of this fact for a unimodal distribution

Theorem 9.3.2 Let f(x) be a unimodal pdf. If the interval $[a, b]$ satisfies

(i) $\int_a^b f(x)dx = 1 - \alpha$
(ii) $f(a) = f(b) > 0$, and
(iii) $a \leq x^* \leq b$ where x^* is a mode of $f(x)$

then $[a, b]$ is the shortest among all intervals that satisfy (i).

Proof (skip details in class; uses mean value theorem to show coverage outside of peak is lower): Let $[a', b']$ be any interval shorter than our interval so $b' - a' < b - a$. We will show that this implies lower coverage $\int_{a'}^{b'} f(x)dx < 1 - \alpha$. We will prove this for $a' \leq a$.

Case $b' \leq a$: $\Rightarrow a' \leq b' \leq a \leq x^*$ and

$$\begin{aligned}
\Rightarrow \int_{a'}^{b'} f(x)dx &\leq f(b')(b' - a') \quad (x \leq b' \leq x^* \Rightarrow f(x) \leq f(b')) \\
&\leq f(a)(b' - a') \quad (b' \leq a \leq x^* \Rightarrow f(b') \leq f(a)) \\
&< f(a)(b - a) \quad (b' - a' < b - a \text{ and } f(a) > 0) \\
&\leq \int_a^b f(x)dx \quad ((ii)(iii) \text{ and unimodality } \Rightarrow f(x) \geq f(a) \text{ for } a \leq x \leq b) \\
&= 1 - \alpha
\end{aligned}$$

Case $b' > a$:

If $b' \geq b \Rightarrow b' - a' \geq b - a$ which is false, so $b' < b$ and more specifically $\Rightarrow a' \leq a < b' < b$.

$$\begin{aligned}\Rightarrow \int_{a'}^{b'} f(x)dx &= \int_{a}^{b} f(x)dx + \left[\int_{a'}^{a} f(x)dx - \int_{b'}^{b} f(x)dx\right] \\ &= (1-\alpha) + \left[\int_{a'}^{a} f(x)dx - \int_{b'}^{b} f(x)dx\right] \\ &= (1-\alpha) + B\end{aligned}$$

Is B negative?

From the unimodality of f, $\Rightarrow a' \leq a < b' < b$, and (ii) and arguments similar to above we have:

$$\int_{a'}^{a} f(x)dx \leq f(a)(a-a') \text{ and } \int_{b'}^{b} f(x)dx \geq f(b)(b-b')$$

Thus,

$$\begin{aligned}B &= \int_{a'}^{a} f(x)dx - \int_{b'}^{b} f(x)dx \\ &\leq f(a)(a-a') - f(b)(b-b') \\ &= f(a)[(a-a') - (b-b')] \quad (f(a) = f(b)) \\ &= f(a)[(b'-a') - (b-a)] \\ &< 0 \quad ((b'-a') < (b-a) \text{ and } f(a) > 0)\end{aligned}$$

Note The form of likelihood regions has optimal construction by this theorem. Recall (re-read example 9.2.3 and discussion)

- The region obtained by inverting the LRT of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ is of the form:

  $$\text{accept } H_0 \text{ if } \frac{\mathcal{L}(\theta_0|\mathbf{x})}{\mathcal{L}(\widehat{\theta}|\mathbf{x})} \leq k(\theta_0)$$

  with the resulting confidence region

  $$\{\theta : \mathcal{L}(\theta|\mathbf{x}) \geq k'(\mathbf{x}, \theta)\},$$

  for some function k' that gives $1-\alpha$ confidence
- In some cases (i.e. Normal and Gamma distribution) the function k' will not depend on θ
- In this case the likelihood region is the set containing θ for which the likelihood is highest
- Now, we know we have the optimal length if $f(a) = f(b)$ for a unimodal distribution

Example 9.3.3 Optimizing expected length For normal intervals of μ with unknown variance based on the pivot $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ we know that the shortest length $1-\alpha$ confidence interval of the form

$$\left[\bar{x} - b\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} - a\frac{s}{\sqrt{n}}\right]$$

has $a = -t_{n-1,\alpha/2}$ and $b = t_{n-1,\alpha/2}$. The interval length is a function of s, with general form

$$length(x) = (b-a)\frac{s}{\sqrt{n}}.$$

We can instead consider the criterion of *expected length* and seek an interval with $1-\alpha$ coverage that minimizes

$$E_\sigma(length(S)) = (b-a)\frac{E_\sigma S}{\sqrt{n}} = (b-a)c(n)\frac{\sigma}{\sqrt{n}},$$

(where the quantity $c(n)$ is a constant dependent only on n) then Theorem 9.3.2 applies and the choice $a = -t_{n-1,\alpha/2}$ and $b = t_{n-1,\alpha/2}$ again gives us an optimal interval.

Example 9.3.4: Shortest pivotal interval (scale parameter) Let $X \sim gamma(k, \beta)$. Then $Y = X/\beta$ is a pivot, with $Y \sim gamma(k, 1)$, so we can get a confidence interval by finding constants a and b to satisfy

$$P(a \leq Y \leq b) = 1-\alpha$$

However, if we apply Theorem 9.3.2 we won't have the shortest confidence interval:

If we choose a and b such that $P(a \leq Y \leq b) = 1-\alpha$ and $f_Y(a) = f_Y(b)$ then the interval on β is of the form:

$$RR = \left\{x : a \leq \frac{x}{\beta} \leq b\right\} \Rightarrow$$

$$CI = \left\{\beta : \frac{x}{b} \leq \beta \leq \frac{x}{a}\right\} \Rightarrow length(CI) = \left(\frac{1}{a} - \frac{1}{b}\right)x$$

which is proportional to $(1/a) - (1/b)$ and not to $b-a$.

We can modify Theorem 9.3.2 to apply here.

Condition (a) in Thm 9.3.2 defines b as a function of a, so call it $b(a)$. We must solve the following constrained minimization problem:

- **Minimize, with respect to a:** $\frac{1}{a} - \frac{1}{b(a)}$
- **Subject to** $\int_a^{b(a)} f_Y(y)dy = 1-\alpha$

Differentiate the first equation with respect to a and setting it equal to 0 yields

$$\frac{db}{da} = \frac{b^2}{a^2}$$

Differentiate the second equation which must equal 0 and substitute this in gives

$$f(b)b^2 = f(a)a^2$$

- equations like these arise in interval estimate of the variance of a normal distribution
- the above equations define not he shortest *overall* interval but the shortest *pivotal* interval, that is, the shortest interval based on the pivot X/β
- this result can be generalized using the Neyman-Pearson Lemma (Exercise 9.43)

3.3.1 Bayesian Intervals

- We say that the confidence interval *covers* the parameter, not that the parameter is *is inside* the interval. This denotes that the random quantity is the interval, not the parameter.
- In a Bayesian setting, the parameter is indeed random and we may say that the parameter is inside an interval with some probability, not 0 or 1.
- All Bayesian claims of coverage are made with respect to the posterior distribution of the parameter
- Bayesian set estimates are referred to as *credible sets* rather than confidence sets.
- If $\pi(\theta|\mathbf{x})$ is the posterior distribution of θ given $\mathbf{X} = \mathbf{x}$, then for any set $A \subset \Theta$, the credible probability of A is:
$$P(\theta \in A|\mathbf{x}) = \int_A \pi(\theta|\mathbf{x})d\theta,$$
and A is a *credible set* for θ. If π is a pmf, we replace integrals with sums.
- Now the parameter is random and the interval is fixed (given the data).

Example 9.2.16 Poisson credible set We wish to construct a credible set for the above Poisson example 9.2.15.

Let $X_1, \ldots, X_n \sim_{iid} Poisson(\lambda)$ and assume λ has a Gamma prior pdf, $\lambda \sim \text{gamma}(a, b)$.

The posterior pdf of λ is then (after applying Bayes' Rule):

$$\pi(\lambda|\sum X = \sum x) = \text{gamma}(a + \sum x, [n + (1/b)]^{-1}$$

We can form a credible set for λ in many ways. One simple way is to split α equally between the upper and lower tails:

$$P(\lambda < \lambda_L(t)|T = t) = P(\lambda > \lambda_U(t)|T = t) = \alpha/2$$

Using the fact that if $Y \sim Gamma(\alpha, \beta)$ then $2Y/\beta \sim \chi^2_{2\alpha}$, it follows that $\frac{2(nb+1)}{b}\lambda \sim \chi^2_{2(a+\sum x_i)}$ assuming a is an integer:

$$\alpha/2 = P(\lambda < \lambda_L(t)|T = t) = P\left(\frac{2\lambda}{b/(bn+1)} < \frac{2\lambda_L(t)}{b/(bn+1)}|T = t\right) = P\left(\chi^2_{2(a+t)} < \frac{2\lambda_L(t)}{b/(bn+1)}t\right)$$

so

$$\frac{2\lambda_L(t)}{b/(bn+1)} = \chi^2_{2(a+t),1-\alpha/2} \Rightarrow \lambda_L(t) = \frac{b}{2(bn+1}\chi^2_{2(a+t),1-\alpha/2}$$

Thus a $1 - \alpha$ credible interval is:

$$\left\{\lambda : \frac{b}{2(bn+1)}\chi^2_{2(a+t),1-\alpha/2} \le \lambda \le \frac{b}{2(bn+1)}\chi^2_{2(a+t),\alpha/2}\right\}$$

If we take $a = b = 1$, the $1 - \alpha$ credible interval becomes:

$$\left(\frac{1}{2(n+1)}\chi^2_{2(1+t),1-\alpha/2}, \frac{1}{2(n+1)}\chi^2_{2(1+t),\alpha/2}\right)$$

while the confidence interval (frequentist) we obtained earlier was:

$$\left[\frac{1}{2n}\chi^2_{2(1+t),1-\alpha/2}, \frac{1}{2n}\chi^2_{2(1+t),\alpha/2}\right]$$

Notes:

- In this case (see Figure 9.2.3), the credible set has somewhat shorter intervals than the confidence interval, and the upper endpoints are closer to 0. This reflects the prior, which is pulling the intervals toward 0.
- We can now ask, what is the probability of getting data ($\mathbf{X}$) for which the credible interval contains the parameter λ? We can show this goes to 0 as $\lambda \to \infty$.
- Likewise, we can ask what is the posterior probability that the confidence interval contains the true parameter? This also goes to 0 as $\sum_{i=1}^{n} X_i \to \infty$, unless $b = 1/n$.
- Using Bayesian criteria for frequentist intervals or frequentist criteria for Bayesian intervals can result in poor calculated performance. That is, Bayesian intervals don't necessarily have good frequentist properties and frequentist intervals don't necessarily have good Bayesian properties.

4 Basic Concepts of Random Samples (C&B Chapter 5)

4.1 Random sample from an infinitely large population

Recall the definition of a random sample (or, *iid* random sample):

Definition The random variables $X_1, \dots, X_n$ are called a *random sample* from the population $f_X(x)$ if

1. $X_i, \dots, X_n$ are mutually independent
2. The marginal pdf (pmf) of each X_i is the same function $f_X(x)$

4.2 Finite population sampling (we won't use this in BSTA 552 unless we have time to talk about bootstrapping)

When we say "random sample" we mean that we are sampling from an infinite population. We can alternatively sample from a *finite* population, say $\{x_1, x_2, \dots, x_N\}$ where $N < \infty$ denotes the size of the population. In this case, we are drawing a sample $X_1^*, X_2^*, \dots, X_n^*$ from the set $\{x_1, x_2, \dots, x_N\}$. There are two options for this type of (re-)sampling:

1. Simple random sample *with replacement*
   - Each value x_i is "replaced" after it is selected (every time you draw a number from a hat you put it back in the hat before drawing the next value)
   - Each sample X_i^* has a discrete uniform distribution with equal probability mass $1/N$ on each of the values $x_1, x_2, \dots, x_N$.
   - The X_i^*'s are mutually independent because the process of choosing each x_i^* is the same, regardless of the values that are chosen for any of the other variables $\Rightarrow X_1^*, X_2^*, \dots, X_n^*$ are *iid*.
   - This type of sampling is the basis of the resampling technique known as *bootstrapping*
2. Simple random sample *without replacement*
   - When sampling from the set $\{x_1, x_2, \dots, x_N\}$, each value x_i is not replaced after it is selected (do not put the number back in the hat before choosing the next value)
   - The X_i^*'s are no longer mutually independent because the probability that X_i^* is equal to some value depends on all the other values selected before it, for instance:

$$P(X_1^* = x_i) = 1/N$$

$$P(X_2^* = x_i | X_1^* = x_i) = 0$$

$$\text{however, if } x_j \neq x_i, \text{ then } P(X_2^* = x_i | X_1^* = x_j) = 1/(N-1).$$

   - However, the X_i^*'s are identically distributed (they have the same marginal distributions, see pg 210 for explanation).
   - If the population size N is "large" then this type of sampling is approximately equal to sampling from an infinite population, and the samples are "nearly independent" in that the conditional distributions are very close to the marginal distributions.

5 Convergence concepts (foundations needed for Chapter 10)

In Chapter 10, we will be discussing convergence of statistics to parameters and also asymptotic distributions of statistics. So first, we define what we mean by "convergence" and "asymptotics".

- Asymptotic theory examines what happens to sample quantities (i.e. statistics) when the sample size approaches infinity.
- There are three main types of convergence: convergence in probability, convergence in distribution, and convergence almost surely. We will only discuss the first two in this course.

5.1 Convergence in probability (consistency)

Definition 5.5.1 Convergence in probability: A sequence of random variables, $Y_1, Y_2, \ldots,$ converges in probability to a random variable Y if, for every small number $\delta > 0$,

$$\lim_{n\to\infty} P(|Y_n - Y| \geq \delta) = 0 \text{ or, equivalently, } \lim_{n\to\infty} P(|Y_n - Y| < \delta) = 1$$

We also define convergence in probability to a constant the same way. A sequence of random variables, $Y_1, Y_2, \ldots,$ converges in probability to a constant value θ if, for every $\delta > 0$,

$$\lim_{n\to\infty} P(|Y_n - \theta| \geq \delta) = 0 \text{ or, equivalently, } \lim_{n\to\infty} P(|Y_n - \theta| < \delta) = 1$$

- For notation, we often write convergence in probability as: $Y_n \to_p Y$ or $Y_n \to_p \theta$
- Convergence in probability means that as n approaches infinity the random variable Y_n becomes arbitrarily close to the random variable Y or the constant θ.
- In Chapter 10, we will be using convergence in probability with a sequence of random variables based on our data: $W_n = W_n(X_1, \ldots, X_n)$ such as $W_n = \sum_{i=1}^n X_i$ or $W_n = \bar{X}_n$. Remember, this is called a *statistic*. We are also usually more concerned with convergence in probability of a statistic to a constant value (as opposed to a random variable).
- Note we have added a subscript n to $\bar{X}$ here since the distribution of $\bar{X}$ depends on n.
- In Chapter 10, we call convergence in probability of a statistic (or a sequence of the same sample quantity such as $\bar{X}_n$) to a parameter θ *consistency* of a statistic.

It is often easiest to prove convergence in probability using:

Chebychev's Inequality (Theorem 3.6.1): Let X be a random variable and let $g(x)$ be a nonnegative function. Then, for any $r > 0$,

$$P(g(X) \geq r) \leq \frac{E[g(X)]}{r^2}.$$

To use this in the above definition of convergence in probability, we first square the difference to get rid of the absolute value and then use Chebychev's Inequality:

$$P(|Y_n - Y| \geq \epsilon) = P((Y_n - Y)^2 \geq \epsilon^2) \leq \frac{E[(Y_n - Y)^2]}{\epsilon^4}.$$

so we see that the probability that the distance $|Y_n - Y|$ or $|Y_n - \theta|$ is arbitrarily large can be bound above by $E[(Y_n - Y)^2]$ or $E[(Y_n - \theta)^2]$. Hence, if $E[(Y_n - Y)^2]$ goes to 0 as $n \to \infty$ then we have convergence in probability to Y. We will see an example of this in the proof of the following theorem.

A special and often used example of convergence in probability is the Weak Law of Large Numbers for $W_n = \bar{X}_n$:

Weak Law of Large Numbers (WLNN, Theorem 5.5.2): Let $X_1, X_2, \ldots$ be *iid* random variables with $EX_i = \mu$ and $\text{Var}X_i = \sigma^2 < \infty$. Define

$$\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$$

Then, for every $\epsilon > 0$:

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| < \epsilon) = 1;$$

that is, $\bar{X}_n$ converges in probability to μ.

Proof of WLLN: We can prove this using Chebychev's inequality.

$$P(|\bar{X}_n - \mu| \geq \epsilon) = P((\bar{X}_n - \mu)^2 \geq \epsilon^2) \leq \frac{E(\bar{X}_n - \mu)^2}{(\epsilon^2)^2} = \frac{\text{Var}\bar{X}_n}{(\epsilon^2)^2} = \frac{\sigma^2}{n\epsilon^4}$$

Since $\frac{\sigma^2}{n\epsilon^4} \to 0$ as $n \to \infty$, $P(|\bar{X}_n - \mu| < \epsilon) = \to 0$, as $n \to \infty$.

Example of convergence in probability to a random variable: An example of a sequence of random variables converging to a random variable (not a constant) is $Y_n = Y + Z_n$ where $Z_n \sim N(\frac{1}{n}, \frac{\sigma^2}{n})$. Then

$$\begin{aligned} P(|Y_n - Y| \geq \epsilon) = P(|Z_n| \geq \epsilon) = P(Z_n^2 \geq \epsilon^2) \\ &\leq \frac{E(Z_n^2)}{\epsilon^4} \text{ (by Chebychev's)} \\ &= \frac{Var(Z_n) + [E(Z_n)]^2}{\epsilon^4} \\ &= \frac{\sigma^2}{n\epsilon} + \frac{1}{n^2\epsilon} \to 0 \, \text{as} \, n \to \infty \end{aligned}$$

So $Y_n \to_p Y$.

Example 5.5.3: Convergence in probability (consistency) of S^2 to σ^2: Let $X_1, X_2, \ldots$ be *iid* random variables with $EX_i = \mu$ and $\text{Var}X_i = \sigma^2 < \infty$. Define

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

then, using Chebychev's Inequality, we have

$$P(|S_n^2 - \sigma^2| \geq \epsilon) \leq \frac{E(S_n^2 - \sigma^2)^2}{\epsilon^4} = \frac{\text{Var}S_n^2}{\epsilon^4}$$

and thus, a sufficient condition that S_n^2 converges in probability to σ^2 is that $\text{Var}S_n^2 \to 0$ as $n \to \infty$. The variance of the sample variance can be shown to be $\frac{2\sigma^4}{n-1}$ which tends toward 0 as $n \to \infty$, so we $n \to \infty$:

$$P(|S_n^2 - \sigma^2| \geq \epsilon) \leq \frac{\text{Var}S_n^2}{\epsilon^4} = \frac{2\sigma^4}{(n-1)\epsilon} \to 0$$

Here is a simulation example showing that S_n^2 converges to σ^2 as $n \to \infty$. We have $X_1, \ldots, X_n \sim N(\mu = 1, \sigma^2 = 4)$ for various n. Note that by simulating a finite number of data sets and calculating S_n^2 on each of those data sets, we are approximating the true underlying distribution of S_n^2. First we can show visually how the S_n^2 values converge to σ^2 and show in a table that the variances (estimates based on 5000 data sets) of S_n^2 converges to 0.

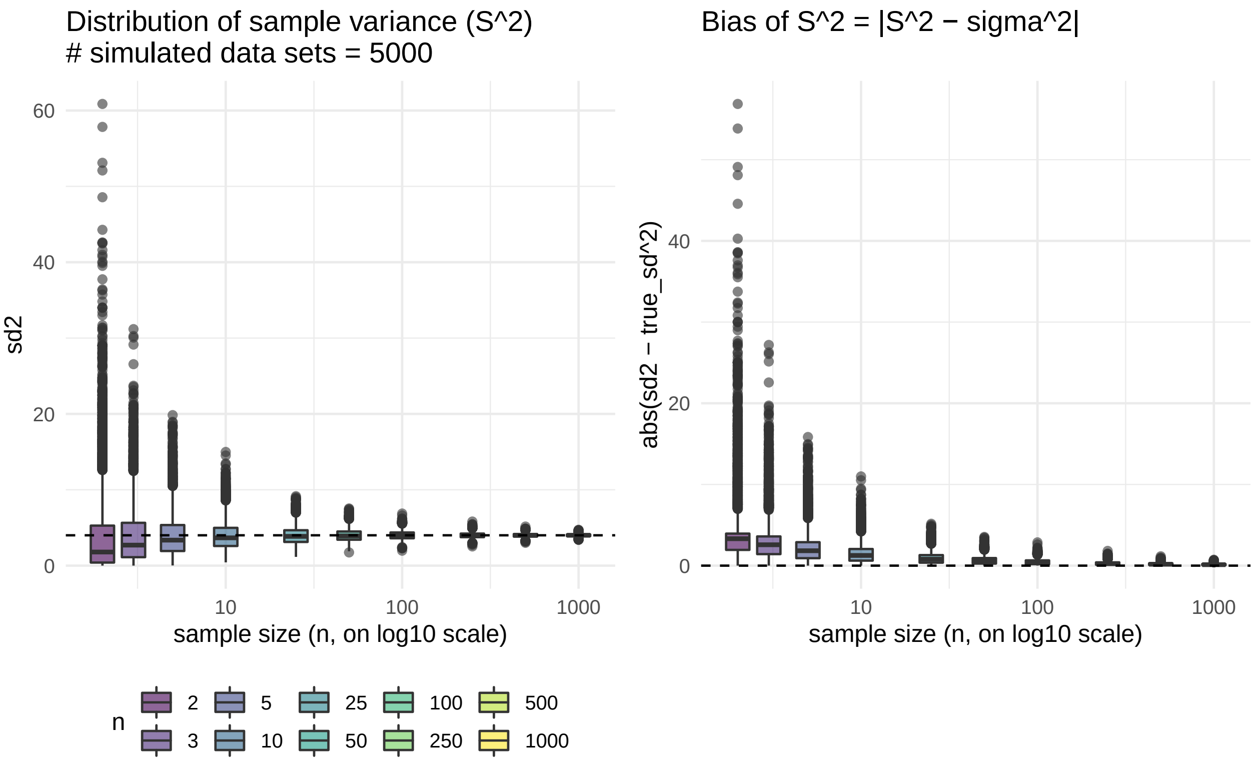
```r
library(tidyverse)
library(patchwork)
true_mean <- 1
true_sd <- 2
nsims <- 5000
nn_all <- c(2, 3, 5, 10, 25, 50, 100, 250, 500, 1000)

set.seed(100)
simdata <- nn_all %>%
  purrr::map_df( ~tibble(x = rnorm(n=.*nsims, true_mean, true_sd),
                         simrep = rep(1:nsims,each=.),
                         nn = .) %>%
  group_by(nn,simrep) %>% summarize(sd2 = sd(x)^2)
  ) %>% ungroup


ggplot2::theme_set(theme_minimal())
p1 <- ggplot(simdata, aes(x=nn, y=sd2, group=nn, fill=factor(nn)))+
  geom_boxplot(alpha=.6)+
  scale_x_log10()+
  xlab("sample size (n, on log10 scale)")+
  geom_hline(yintercept =true_sd^2, lty=2)+
  ggtitle(glue::glue("Distribution of sample variance (S^2)\n# simulated data sets = {nsims}"))
  theme(legend.position="bottom")+
  scale_fill_viridis_d(name="n")

p2 <- ggplot(simdata, aes(x=nn, y=abs(sd2-true_sd^2), group=nn, fill=factor(nn)))+
  geom_boxplot(alpha=.6)+
  scale_x_log10()+
  xlab("sample size (n, on log10 scale)")+
  geom_hline(yintercept =0, lty=2)+
  ggtitle("Bias of S^2 = |S^2 - sigma^2|\n ")+
  scale_fill_viridis_d(guide=FALSE)

p1 + p2
```



```
# calculate variances of S^2
simdata_vars <- simdata %>% group_by(nn) %>% summarize(var(sd2))
knitr::kable(simdata_vars, caption = "Estimate of var(S_n^2) for various n")
```

Table 1: Estimate of var(S_n^2) for various n

nn	var(sd2)
2	33.4532308
3	15.4419739
5	8.0656179
10	3.6600892
25	1.3242583
50	0.6578749
100	0.3222768
250	0.1288456
500	0.0637904

nn	var(sd2)
1000	0.0316714

We also know that continuous functions of a sequence of random variables converge in probability if the sequence of random variables itself converges in probability:

Continuous mapping theorem: (Theorem 5.5.4 Convergence in probability of a function of a sequence of random variables) Suppose that $Y_1, Y_2, \ldots$ converges in probability to a random variable Y (or constant θ) and that h is a continuous function. Then $h(Y_1), h(Y_2), \ldots$ converges in probability to $h(Y)$ (or $h(\theta)$).

For example, since $S^2 \rightarrow_p \sigma^2$ then $\sqrt{S^2} \rightarrow_p \sqrt{\sigma^2}$, or $S \rightarrow_p \sigma$.

5.2 Convergence in distribution

Convergence in distribution examines what happens to the distribution function (specifically, the CDF) of a sequence of random variables as $n \rightarrow \infty$.

Definition 5.5.10 Convergence in distribution A sequence of random variables $Y_1, Y_2, \ldots,$ *converges in distribution* to a random variable Y if

$$\lim_{n \to \infty} F_{Y_n}(y) = F_Y(y)$$

at all points y where $F_Y(y)$ is continuous.

Example 5.5.11 Maximum of uniforms If $X_1, X_2, \ldots$ are *iid* uniform(0,1), what happens to $X_{(n)}$ as $n \rightarrow \infty$? We can show $X_{(n)} \rightarrow_p 1$:

$$\begin{aligned}
P(|X_{(n)} - 1| \geq \epsilon) &= P(X_{(n)} \leq 1 - \epsilon) + P(X_{(n)} \geq 1 + \epsilon) \\
&= P(X_{(n)} \leq 1 - \epsilon) + 0 \\
&= \prod_{i=1}^{n} P(X_i \leq 1 - \epsilon) \\
&= [P(X_i \leq 1 - \epsilon)]^n = (1 - \epsilon)^n \rightarrow 0
\end{aligned}$$

Now we can show $Y_n = n(1 - X_{(n)})$ converges in distribution to $Y \sim$ Exponential(1) random variable. In order to show this, we need to show that the limit of the cdf of $n(1 - X_{(n)})$ is the cdf of an Exponential(1) distribution.

In the above proof of convergence in probability, let $\epsilon = t/n$. Now,

$$P(X_{(n)} \leq 1 - \epsilon) = P(X_{(n)} \leq 1 - t/n) = (1 - t/n)^n \rightarrow e^{-t}$$

The limit is true by the definition of the exponential function. Upon rearranging, we have

$$F_{Y_n}(t) = P(n(1 - X_{(n)}) \leq t) = P(X_{(n)} \leq 1 - t/n) \rightarrow e^{-t} = F_y(t)$$

and hence $Y_n \to_d Y$.

We can simulate this data and show convergence of the (approximate) distribution of the simulated data. First let's show convergence in probability of $X_{(n)} \to_p 1$.

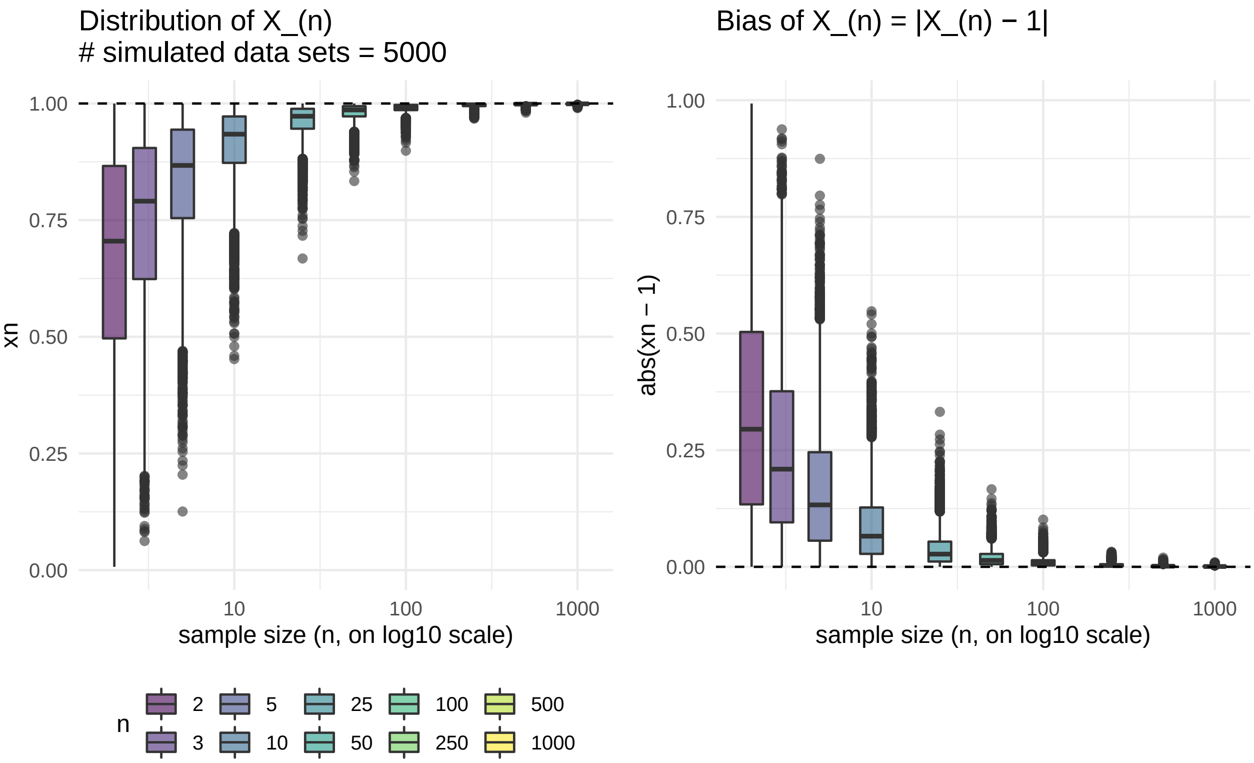
```r
nsims <- 5000
nn_all <- c(2, 3, 5, 10, 25, 50, 100, 250, 500, 1000)

set.seed(100)
simdata <- nn_all %>%
  purrr::map_df( ~tibble(x = runif(n=.*nsims, 0, 1),
                         simrep = rep(1:nsims,each=.),
                         nn = .) %>%
  group_by(nn,simrep) %>%
    summarize(xn = max(x))
  ) %>% ungroup %>%
  mutate(yn = nn*(1-xn))

p1 <- ggplot(simdata, aes(x=nn, y=xn, group=nn, fill=factor(nn)))+
  geom_boxplot(alpha=.6)+
  scale_x_log10()+
  xlab("sample size (n, on log10 scale)")+
  geom_hline(yintercept =1, lty=2)+
  ggtitle(glue::glue("Distribution of X_(n)\n# simulated data sets = {nsims}"))+
  theme(legend.position="bottom")+
  scale_fill_viridis_d(name="n")

p2 <- ggplot(simdata, aes(x=nn, y=abs(xn-1), group=nn, fill=factor(nn)))+
  geom_boxplot(alpha=.6)+
  scale_x_log10()+
  xlab("sample size (n, on log10 scale)")+
  geom_hline(yintercept =0, lty=2)+
  ggtitle("Bias of X_(n) = |X_(n) - 1|\n ")+
  scale_fill_viridis_d(guide=FALSE)

p1 + p2
```



```r
# calculate variances of X_(n)
simdata_vars <- simdata %>% group_by(nn) %>% summarize(var(xn))
knitr::kable(simdata_vars, caption = "Estimate of var(X_(n)) for various n")
```

Table 2: Estimate of var(X_(n)) for various n

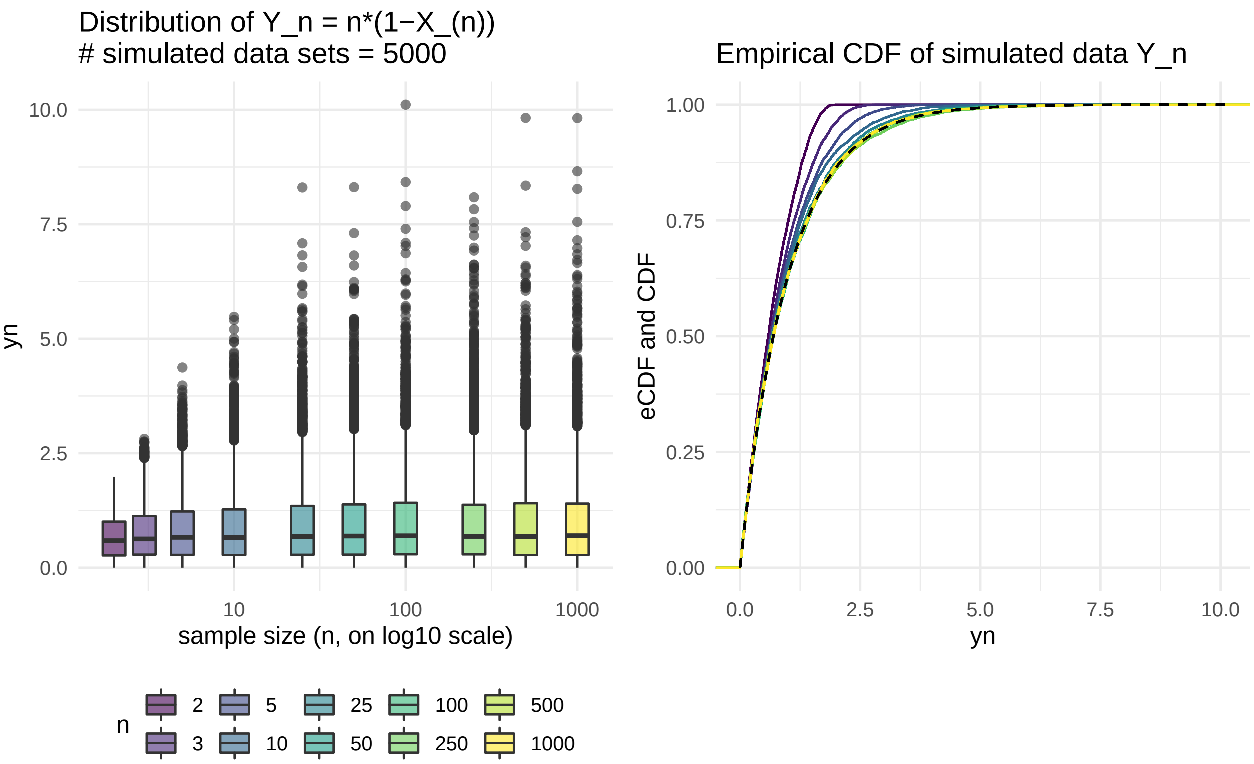
nn	var(xn)
2	0.0549248
3	0.0376370
5	0.0198240
10	0.0067927
25	0.0013699
50	0.0003746
100	0.0000989
250	0.0000171
500	0.0000040
1000	0.0000010

Now, we show convergence in distribution of $Y_n = n(1 - X_{(n)})$. Note first that Y_n does not appear to converge in probability to a constant since the variance does not get smaller as $n \to \infty$ (left panel). [Note the median looks like it is stabilizing, and in fact it is converging to the median of Exp(1) which is $\ln(2) = 0.693\ldots$]. However, if we look at the approximate distribution of Y_n compared to an Exp(1) distribution, we can see convergence of the CDFs to the CDF of Exp(1) represented by the dashed line (right panel).

```r
p1 <- ggplot(simdata, aes(x=nn, y=yn, group=nn, fill=factor(nn)))+
  geom_boxplot(alpha=.6)+
  scale_x_log10()+
  xlab("sample size (n, on log10 scale)")+
  ggtitle(glue::glue("Distribution of Y_n = n*(1-X_(n))\n# simulated data sets = {nsims}"))+
  theme(legend.position="bottom")+
  scale_fill_viridis_d(name="n")

p2 <- ggplot(simdata, aes(x=yn, group=nn, color=factor(nn)))+
  stat_ecdf()+
  ggtitle("Empirical CDF of simulated data Y_n")+
  ylab("eCDF and CDF")+
  stat_function(fun=pexp, color="black", lty=2)+
  scale_color_viridis_d(guide=FALSE)

p1+p2
```



5.2.1 Important theorems about convergence in distribution

Convergence in probability is "stronger" than convergence in distribution since $\rightarrow_p$ implies $\rightarrow_d$.

Theorem 5.5.12 Convergence in probability implies converge in distribution: If the sequence of random variables, $Y_1, Y_2, \ldots,$ converges in probability to a random variable Y, the sequence also converges in distribution to Y.

Convergence in distribution to a constant has a special property that it is equivalent to convergence in probability.

Theorem 5.5.13 Convergence in distribution to constant = convergence in probability to constant: The sequence of random variables $Y_1, Y_2, \ldots,$ converges in probability to a constant μ if and only if the sequence also converges in distribution to μ. That is, the statement

$$P(|Y_n - \mu| > \delta) \rightarrow 0 \text{ for every } \delta > 0$$

is equivalent to

$$P(Y_n \leq y) \rightarrow \begin{cases} 0, & \text{if } y < \mu \\ 1, & \text{if } y > \mu. \end{cases}$$

An important and famous example of convergence in distribution is the convergence of the distribution of the sample mean, called the Central Limit Theorem. One version of the CLT is as follows:

Central Limit Theorem (Theorem 5.5.14) Let $X_1, X_2, \ldots$ be a sequence of *iid* random variables whose moment generating functions (mgfs) exist in a neighborhood of 0 (that is, $M_{X_i}(t)$ exists for $|t| < h$, for some positive h). Let $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 > 0$. (Both μ and σ^2 are finite since the mgf exists.) Define $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$. Let $G_n(x)$ denote the cdf of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$. Then, for any $x, -\infty < x < \infty$,

$$\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy;$$

that is, $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ has a limiting standard normal distribution:

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \rightarrow_d \mathcal{N}(0,1)$$

See the proof in C&B (uses Taylor series expansion of the moment generating function).

A very useful theorem for proving convergence in distribution when you have a sum or product of random variables is Slutsky's Theorem:

Slutsky's Theorem (Theorem 5.5.17): If $X_n \xrightarrow{d} X$ in distribution and $Y_n \xrightarrow{p} a$, with a constant, in probability, then:

a. $Y_n X_n \xrightarrow{d} aX$ (in distribution)
b. $X_n + Y_n \xrightarrow{d} X + a$ (in distribution)

- Note that Y_n must converge in probability to a constant, not a random variable. Otherwise, these relationships do not always hold.

- Also, if $X_n \to_p X$ in probability, the above theorem still holds since this implies that $X_n \to_d X$ in distribution.

Example 5.5.18 Normal approximation with estimated variance: Suppose that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \to_d \mathcal{N}(0,1)$$

but the value of σ is unknown. We have seen in Example 5.5.3 that, if $\lim_{n\to\infty} \text{Var}S_n^2 = 0$, then $S_n^2 \to \sigma^2$ in probability. We can show (Exercise 5.32) that $\sigma/S_n \to 1$ in probability. Therefore, by Slutsky's Theorem

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} = \frac{\sigma}{S_n}\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \to_d \mathcal{N}(0,1).$$

Another very useful theorem about convergence in distribution is the Delta Method. It is used when we need to determine the asymptotic distribution of a function of a statistic. You can think of this as "a generalized CLT" since it can be used to show convergence of a function of $\bar{X}$.

Delta Method (Theorem 5.5.24) Let Y_n be a sequence of random variables that satisfies

$$\sqrt{n}(Y_n - \theta) \xrightarrow{d} \mathcal{N}(0,\sigma^2).$$

For a given function g and a specific value of θ, assuming $g'(\theta)$ exists and is not 0, then:

$$\sqrt{n}\,[g(Y_n) - g(\theta)] \xrightarrow{d} \mathcal{N}(0,\sigma^2[g'(\theta)]^2).$$

Proof: The Taylor expansion of $g(Y_n)$ around $Y_n = \theta$ is

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + \text{Remainder}$$

where the remainder $\to 0$ as $Y_n \to \theta$. Since $Y_n \to \theta$ in probability it follows that the remainder (call it Z_n) $\to 0$ in probability. So, by applying Slutsky's Theorem

$$\sqrt{n}[g(Y_n) - g(\theta)] = g'(\theta)\sqrt{n}(Y_n - \theta) + Z_n \to_d g'(\theta)\mathcal{N}(0,\sigma^2) + 0 = \mathcal{N}(0,\sigma^2[g'(\theta)]^2)$$

and the result follows.

Example 5.5.25: Suppose X_i is a random variable with $E_\mu X_i = \mu \neq 0$. Suppose we want to find the distribution of $\frac{1}{\bar{X}}$. By the Delta Method we have $g'(\mu) = -\mu^{-2}$ and

$$\sqrt{n}\left(\frac{1}{\bar{X}} - \frac{1}{\mu}\right) \to \mathcal{N}\left(0, \left(\frac{1}{\mu}\right)^4 \text{Var}_\mu X\right)$$

However, we may not know the variance of X, so we need to estimate it with S^2. Also, we need to estimate μ with $\bar{X}$. We can estimate the whole variance:

$$\widehat{\text{Var}}\left(\frac{1}{\bar{X}}\right) \approx \left(\frac{1}{\bar{X}}\right)^4 S^2.$$

Since both $\bar{X}$ and S^2 are consistent estimators of μ and σ^2, we can apply Slutsky's Theorem to conclude that for $\mu \neq 0$,

$$\frac{\sqrt{n}\left(\frac{1}{\bar{X}} - \frac{1}{\mu}\right)}{\left(\frac{1}{\bar{X}}\right)^2 S} \to_d \mathcal{N}(0,1)$$

in distribution.

Sometimes the derivative of g is 0 at θ so we must use the second-order Delta Method:

Second-order Delta Method (Theorem 5.5.26) Let Y_n be a sequence of random variables that satisfies

$$\sqrt{n}(Y_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

For a given function g and a specific value of θ, suppose that $g'(\theta) = 0$ and $g''(\theta)$ exists and is not 0. Then

$$n[g(Y_n) - g(\theta)] \to \sigma^2 \frac{g''(\theta)}{2} \chi_1^2$$

That is, the asymptotic distribution is a chi-square random variable. [This is also proven using Taylor Series expansions but out to the second degree polynomial, and Slutsky's theorem.]

6 Asymptotic Evaluations (Chapter 10, with some review of Chapters 5 and 7)

- So far we have considered *finite-sample criteria* for inference about a parameter θ. That is, the distributions of test statistics were valid for finite sample sizes.
- Asymptotics are concerned with the properties of estimators, random variables, and hypothesis tests as the sample size(s) increases.
- “asymptotics uncover the most fundamental properties of a procedure and give us a very powerful and general evaluation tool”–C&B
- We are concerned with a sequence of estimators $W_n = W_n(X_1, \dots, X_n)$. Often $W_n = \widehat{\theta}_n$ the MLE of θ.

6.1 Consistency $\sim$ relates to the expectation of an estimator

- As $n \to \infty$, an estimator should converge to the “correct" value = the parameter of interest.
- Equivalently, bias should go to 0; we want our estimator to be "asymptotically unbiased".
- We would not want an inconsistent estimator, since then we are tending towards bias no matter how large of a sample (even the entire population) we obtain.

Definition 10.1.1 Consistency (convergence in probability to a constant) A sequence of estimators $W_n = W_n(X_1, \dots, X_n)$ is a *consistent sequence of estimators* of the parameter θ if, $W_n \xrightarrow{p} \theta$.

- Previously we saw that $\bar{X}_n \xrightarrow{p} \mu = E(X_i)$ by the WLLN and $S_n^2 \xrightarrow{p} \sigma^2 = Var(X_i)$ so these are examples of consistent estimators.

6.1.1 Mean Square Error $\to 0$ implies consistency

In order to prove consistency, we can also examine the behavior of the finite sample variance and bias of our estimator. Together, this is the mean square error.

- For an estimator $\widehat{\theta}$ of θ we wish to performance or ‘goodness’ of the estimator.
- Hence we estimate the deviation from $\widehat{\theta}$ to the true value.
- The *absolute error* $|\widehat{\theta} - \theta|$ measures this.
- However, the square of this $(\widehat{\theta} - \theta)^2$ also measures this deviation but has nicer mathematical properties

Definition The *mean square error (MSE)* of an estimator W_n of the parameter θ is the function $E(W_n - \theta)^2$. We can denote this as $MSE_\theta(W_n)$. This is also called the quadratic loss function.

An important mathematical property of MSE is (from 7.3.1):

$$MSE_\theta(W_n) = E_\theta[(W_n - \theta)^2] = E_\theta[(W_n - E_\theta(W_n))^2] + [E_\theta(W_n) - \theta]^2 = Var_\theta W_n + [Bias_\theta W_n]^2$$

- The MSE has two components: the variability of an estimator (precision) and the bias (accuracy)

- We need to find estimators that control both variance and bias.
- For an unbiased estimator we only need to control variance since then $MSE_{\widehat{\theta}} = Var(\widehat{\theta})$

In general when estimating a parameter θ via W_n, if $Var_\theta(W_n) \to 0$ and $Bias_\theta(W_n) \to 0$ as $n \to \infty$, then the estimator is consistent for θ. This can be seen since by using Chebychev's Inequality (Theorem 3.6.1) to link the definition of convergence in probability to the MSE:

$$P_\theta(|W_n - \theta| \geq \delta) \leq \frac{E_\theta[(W_n - \theta)^2]}{\delta^4}$$

so if, for every $\theta \in \Theta$,

$$\lim_{n\to\infty} E_\theta[(W_n - \theta)^2] = \lim_{n\to\infty} MSE_\theta(W_n) = 0,$$

then the sequence of estimators is consistent for θ. Furthermore, by (7.3.1),

$$MSE_\theta(W_n) = E_\theta[(W_n - \theta)^2] = Var_\theta W_n + [Bias_\theta W_n]^2$$

Combining these facts, we have the following theorem:

Theorem 10.1.3 If W_n is a sequence of estimators of a parameter θ satisfying

1. $\lim_{n\to\infty} Var_\theta W_n = 0$,
2. $\lim_{n\to\infty} Bias_\theta W_n = 0$,

for every $\theta \in \Theta$, then W_n is a consistent sequence of estimators of θ.

In other words, if $MSE_\theta(W_n) \to 0$, then W_n is consistent for θ.

Example general variance Let $X_1, X_2, \ldots X_n$ be iid $\mathcal{N}(\mu, \sigma^2)$ with known $\sigma < \infty$. Then $W_n = \bar{X}_n$ is an estimator of μ.

1. $\lim_{n\to\infty} Var_\theta \bar{X} = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$,
2. $\lim_{n\to\infty} Bias_\theta \bar{X} = \lim_{n\to\infty} E(\bar{X}) - \mu = \mu - \mu = 0$,

So by Thm 10.1.3, $\bar{X}_n$ is a consistent estimator of μ.

6.1.2 Linear combinations of consistent estimators

We can also construct consistent estimators from other consistent estimators:

Theorem 10.1.5 Let W_n be a consistent sequence of estimators of a parameter θ. Let $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$ be sequences of constants satisfying

1. $\lim_{n\to\infty} a_n = 1$,
2. $\lim_{n\to\infty} b_n = 0$.

Then the sequence $U_n = a_n W_n + b_n$ is a consistent sequence of estimators of θ.

Proof: Use Slutsky's theorem from chapter 5.

Example: Consistency of Normal variance estimates Let $X_1, \ldots, X_n$ be iid $\mathcal{N}(\mu, \sigma^2)$. Let $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$. We have seen that $E_{\mu,\sigma^2}(S^2) = \sigma^2$ and $Var_{\mu,\sigma^2}(S^2) = 2\frac{\sigma^4}{n-1}$. So, S^2 is a consistent estimator of σ^2.

Since $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n}{n-1}S^2$, $\widehat{\sigma}^2$ is also a consistent estimator of σ^2.

We could have also proven this by showing that $\widehat{\sigma}^2$ is asymptotically unbiased (even though it has bias in finite samples, the bias goes to 0 as $n \to \infty$) and that the variance of $\widehat{\sigma}^2$ goes to 0.

6.1.3 Consistency of MLEs

A very important result!

Theorem 10.1.6 Consistency of MLEs Let $X_1, X_2, \ldots$ be iid $f(x|\theta)$, and let $L(\theta|\mathbf{x}) = \prod_{i=1}^n f(x_i|\theta)$ be the likelihood function. Let $\widehat{\theta}$ denote the MLE of θ. Let $\tau(\theta)$ be a continuous function of θ. Under certain regularity conditions on $f(x|\theta)$ and, hence, $L(\theta|\mathbf{x})$, for every $\delta > 0$ and every $\theta \in \Theta$,

$$\lim_{n\to\infty} P_\theta(|\tau(\widehat{\theta}) - \tau(\theta)| \geq \delta) = 0.$$

That is, $\tau(\widehat{\theta})$ is a consistent estimator of $\tau(\theta)$.

Proof sketch: Show that $\frac{1}{n}\log \mathcal{L}(\widehat{\theta}|\mathbf{x})$ converges almost surely (a stronger convergence than convergence in probability) to $E_\theta(\log f(X|\theta))$ for every $\theta \in \Theta$. This (under some conditions) implies that $\widehat{\theta} \xrightarrow{p} \theta$ and, hence $\tau(\widehat{\theta}) \xrightarrow{p} \tau(\theta)$.

For regularity conditions (related to identifiability, differentiable f, etc) see Miscellanea 10.6.2.

- There are extensions to non-iid settings, for example regression.
- So, under regularity conditions, MLE's get closer and closer to the parameters they are estimating as the sample size gets large.

6.2 Efficiency $\sim$ relates to the variance of the estimator

- Consistency is concerned with asymptotic accuracy. We should also be concerned with the asymptotic variance of an estimator.
- We'd also like the MLE estimator to have small variance for large samples and to have a limiting distribution that we can use for obtaining tests and confidence intervals.
- In many cases for an estimator W_n, $\text{Var}(W_n) \to 0$ as $n \to \infty$ since it is a consistent estimator. So we need to evaluate the variance of $k_n W_n$ where k_n is some normalizing constant to force the variance to a limit that is not 0.

6.2.1 Efficient Estimator (finite sample property)

Recall **Fisher's Information** and the three equivalent definitions:

$$I(\theta) = I_1(\theta) = E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_X(x|\theta)\right)^2\right] = Var_\theta\left[\frac{\partial}{\partial\theta}\log f_X(x|\theta)\right] = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f_X(x|\theta)\right]$$

We use the notation $I_n(\theta)$ to denote the information number (expected information) based on a sample of n observations.

$$I_n(\theta) = -E_\theta \left[\frac{\partial^2}{\partial\theta^2} \sum_{i=1}^{n} \log f_X(x_i|\theta) \right] = -\sum_{i=1}^{n} E_\theta \left[\frac{\partial^2}{\partial\theta^2} \log f_X(x_i|\theta) \right] = nI_1(\theta)$$

We use Fisher's Information to bound the variance of an estimator with Cramer-Rao's Inequality (a.k.a the information inequality):

$$Var(\widehat{\theta}) \geq \frac{[\frac{\partial}{\partial\theta} E_\theta(\widehat{\theta})]^2}{nI_1(\theta)}$$

As the information number increases, we can bound the variance with a smaller number.

Definition: When the equality holds, that is, $Var(\widehat{\theta}) = \frac{[\frac{\partial}{\partial\theta} E_\theta(\widehat{\theta})]^2}{nI(\theta)}$, the estimator $\widehat{\theta}$ is said to be an *efficient estimator* of its expectation $E_\theta(\widehat{\theta})$. If $E_\theta(\widehat{\theta}) = \theta$ then $\widehat{\theta}$ is an efficient estimator of θ. If an estimator is unbiased and its variance reaches the CR lower bound, then it is the minimum variance unbiased estimator (MVUE).

6.2.2 Asymptotic variance

- In many cases for an estimator W_n, $\text{Var}(W_n) \to 0$ as $n \to \infty$ since it is a consistent estimator. So we need to evaluate the variance of $k_n W_n$ where k_n is some normalizing constant (function of n) to force the variance to a limit that is not 0.
- In particular, we want to choose $\{k_n\}$ such that we have an asymptotically normal distribution with a particular asymptotic variance.
- Then we can compare variances of the asymptotic distributions for different estimators to see which is smallest.

Definition 10.1.9 For an estimator W_n, suppose that $k_n(W_n - \tau(\theta)) \xrightarrow{d} \mathcal{N}(0, \sigma^2))$ in distribution. That is,

$$\lim_{n\to\infty} \text{cdf of } k_n(W_n - \tau(\theta)) = \text{ cdf of } \mathcal{N}(0, \sigma^2).$$

The parameter σ^2 is called the *asymptotic variance* of W_n.

- If two estimators have the same asymptotic distribution and the same asymptotic mean but different asymptotic variances, we prefer the estimator with the smaller variance.

Definition 10.1.11 Let W_n be a sequence of estimators based on $X_1, \ldots, X_n \sim_{iid} f(x|\theta)$. W_n is *asymptotically efficient* for a parameter $\tau(\theta)$ if $\sqrt{n}(W_n - \tau(\theta)) \to \mathcal{N}(0, v[\tau(\theta)])$ in distribution where $v[\tau(\theta)] > 0$ and

$$v[\tau(\theta)] = \frac{[\tau'(\theta)]^2}{E_\theta((\frac{\partial}{\partial\theta} \log f(X|\theta))^2)} = \frac{[\tau'(\theta)]^2}{I_1(\theta)}.$$

That is, the asymptotic variance of W_n achieves the Cramer-Rao Lower Bound for unbiased estimators of $\tau(\theta)$.

When $\tau(\theta) = \theta$, then $\tau'(\theta) = 1$ so

$$v[\theta] = \frac{1}{I_1(\theta)}.$$

- Note, because we are multiplying by $\sqrt{n}$ we don't need the n in the CRLB– we use the density for a *single observation* in the *iid* case. That is why the quantity in the denominator is based on the pdf (pmf) of a single observation.
- Theorem 10.1.6 stated that MLEs are consistent (under general conditions). Under somewhat stronger regularity conditions, we have similar result for asymptotic efficiency.

Theorem 10.1.12 Asymptotic efficiency of MLEs: Let $X_1, X_2, \ldots$ be $\sim_{iid} f(x|\theta)$, let $\widehat{\theta}$ denote the MLE of θ, and let $\tau(\theta)$ be a continuous function of θ. Under the regularity conditions in Misc. 10.6.2 on $f(x|\theta)$ and hence, $\mathcal{L}(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta)$,

$$\sqrt{n}[\tau(\widehat{\theta}) - \tau(\theta)] \xrightarrow{d} \mathcal{N}[0, v[\tau(\theta)]$$

where $v[\tau(\theta)]$ is the Cramer-Rao Lower Bound. That is, $\tau(\widehat{\theta})$ is a consistent and asymptotically efficient estimator of $\tau(\theta)$. (!)

So, the asymptotic variance of the MLE achieves the smallest possible variance among unbiased estimators of $\tau(\theta)$.

Proof: We first outline the proof showing the case that $\tau(\theta) = \theta$ to show that $\widehat{\theta}$ is asymptotically efficient.

Let $\ell(\theta|\mathbf{x}) = \sum_{i=1}^{n} \log f(\mathbf{x}_i|\theta)$ denote the log-likelihood function and denote its derivatives by $\ell', \ell'', \ldots$. Obtain a Taylor's expansion of the first derivative about the true value θ_0:

$$\ell'(\theta|\mathbf{x}) = \ell'(\theta_0|\mathbf{x}) + (\theta - \theta_0)\ell''(\theta_0|\mathbf{x}) + \mathit{Remainder}$$

where *Remainder* represents the remaining terms that converge to 0 (specifically, they have the property $\mathit{Remainder}/n \to 0$ as $n \to \infty$ under the regularity conditions).

Substitute $\widehat{\theta}$ for θ:

$$\ell'(\widehat{\theta}|\mathbf{x}) = \ell'(\theta_0|\mathbf{x}) + (\widehat{\theta} - \theta_0)\ell''(\theta_0|\mathbf{x}) + \mathit{Remainder}$$

The left-hand side $\ell'(\widehat{\theta}|\mathbf{x})$ is equal to 0 since $\widehat{\theta}$ is an MLE and we find the MLE by setting the derivative of the log-likelihood to 0. Rearranging we have:

$$\begin{aligned}
(\widehat{\theta} - \theta_0) &= -\frac{\ell'(\theta_0|\mathbf{x})}{\ell(\theta_0|\mathbf{x}) + \mathit{Remainder}} \Leftrightarrow \\
\sqrt{n}(\widehat{\theta} - \theta_0) &= -\sqrt{n}\frac{\ell'(\theta_0|\mathbf{x})}{\ell''(\theta_0|\mathbf{x}) + \mathit{Remainder}} \\
&= -\frac{\ell'(\theta_0|\mathbf{x})/\sqrt{n}}{\ell''(\theta_0|\mathbf{x})/n + \mathit{Remainder}/n}
\end{aligned}$$

Since the remainder term $\mathit{Remainder}/n \xrightarrow{p} 0$, to find the asymptotic distribution of $\sqrt{n}(\widehat{\theta} - \theta_0)$ we need to find the asymptotic distribution of

$$\frac{\ell'(\theta_0|\mathbf{x})/\sqrt{n}}{-\ell''(\theta_0|\mathbf{x})/n}$$

Let $I_1(\theta_0) = 1/v(\theta)$ denote the information number for one observation.

First consider the numerator:

Now, since $X_1, \ldots, X_n$ are *iid* and $E(\ell'(\theta_0|x_i)) = 0$, by the Central Limit Theorem:

$$\frac{\ell'(\theta_0|\mathbf{x})}{\sqrt{n}} = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\ell'(\theta_0|x_i)\right] \xrightarrow{d} \mathcal{N}(0, Var_{\theta_0}[\ell'(\theta_0|x_i)])$$

where

$$Var_{\theta_0}[\ell'(\theta_0|x_i)] = E_{\theta_0}[-\ell''(\theta_0|x_i)] = I_1(\theta_0) = 1/v(\theta) \text{ (definition of Fisher's Information)}$$

Now consider the denominator:

Also, by the Weak Law of Large Numbers:

$$-\frac{1}{n}\ell''(\theta_0|\mathbf{X}) = -\frac{1}{n}\sum_{i=1}^{n}\ell''(\theta_0|x_i) \xrightarrow{p} I_1(\theta_0).$$

Num/Denom:

So, by Slutsky's Theorem:

$$\sqrt{n}(\widehat{\theta} - \theta_0) = \left[-\frac{1}{n}\ell''(\theta_0|\mathbf{X})\right]^{-1}\left[\frac{\ell'(\theta_0|\mathbf{x})}{\sqrt{n}}\right] \xrightarrow{d} [I_1(\theta_0)]^{-1}\,\mathcal{N}(0, I_1(\theta_0)) = \mathcal{N}(0, 1/I_1(\theta_0)) = \mathcal{N}(0, v(\theta))$$

To extend to a general $\tau(\theta)$, the Delta Method says that

$$\sqrt{n}(\tau(\theta) - \tau(\theta)) \xrightarrow{d} \mathcal{N}(0, v(\theta)[\tau'(\theta)]^2)$$

and in fact $v(\theta)[\tau'(\theta)]^2 = [\tau'(\theta)]^2/I(\theta_0)$ is the CR lower bound for $\tau(\theta)$.

- Note on regularity conditions: For *iid* samples from one parameter *regular* exponential families (parameter space contains an open interval), the MLE (obtained as a solution to likelihood equations) is asymptotically efficient. For multiparameter exponential families the results also hold provided the likelihood equations for the MLE have a solution.

Example 10.1.13 Asymptotic normality and consistency The above theorem shows that typically MLEs are efficient and consistent. However, this is somewhat redundant, as efficiency is defined only when the estimator is asymptotically normal and, as we will illustrate, *asymptotic normality implies consistency.*

Suppose that

$$\sqrt{n}(W_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

then

$$\sqrt{n}\frac{W_n - \theta}{\sigma} \xrightarrow{d} \frac{1}{\sigma}\mathcal{N}(0, \sigma^2) = \mathcal{N}(0, 1)$$

From Slutksy's Theorem:

$$W_n - \mu = \frac{\sigma}{\sqrt{n}}\left(\sqrt{n}\frac{W_n - \theta}{\sigma}\right) \to \lim_{n\to\infty} \frac{\sigma}{\sqrt{n}}\mathcal{N}(0,1) = 0.$$

so $W_n - \mu \overset{d}{\to} 0$ which implies convergence in probability to 0 (Thm 5.5.13 above). Hence, W_n is a consistent estimator of μ.

Example Normal MLE Let $X_1, \ldots, X_n \sim_{iid} \mathcal{N}(\mu, \sigma^2), \sigma^2$ known. Then the MLE of μ is $\bar{X}$ and $\sqrt{n}(\bar{X} - \mu) \sim \mathcal{N}(0, \sigma^2)$. This is the exact distribution. In this case, this is also the asymptotic distribution. This result follows from the CLT but also the above theorem since

$$\begin{aligned}
I_1(\mu) &= E_\mu\left(-\frac{\partial^2}{\partial\mu^2}\log f(x|\mu)\right) \\
&= -E_\mu\left(\frac{\partial^2}{\partial^2\mu} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}\right) \\
&= -E_\mu\left(\frac{\partial}{\partial\mu}\frac{(x-\mu)}{\sigma^2}\right) \\
&= E_\mu\frac{1}{\sigma^2} \\
&= \frac{1}{\sigma^2}
\end{aligned}$$

so

$$v(\mu) = 1/I_1(\mu) = \sigma^2$$

and by Thm 10.1.12

$$\sqrt{n}(\bar{X} - \mu) \to_d \mathcal{N}(0, \sigma^2)$$

Example 10.1.14 Binomial variance Let $X_1, \ldots, X_n \sim_{iid} Bernoulli(p)$. We know that the MLE for p is $\widehat{p} = \bar{X}$. Since $E(\bar{X}) = p$ and $Var(\bar{X}) = \frac{1}{n}p(1-p)$, and $\bar{X}$ is a sample mean, the CLT tells us

$$\sqrt{n}(\widehat{p} - p) \overset{d}{\to} \mathcal{N}(0, p(1-p)).$$

However, we could have obtained the variance by $v(p) = 1/I(p)$ and the theorem above.

$$f_X(x|p) = p^x(1-p)^{(1-x)}$$

$$\log f_X(x|\theta) = x\log p + (1-x)\log(1-p)$$

$$\frac{\partial}{\partial\theta^2}\log f_X(x|\theta) = \frac{x}{p} - \frac{1-x}{1-p}$$

$$\frac{\partial^2}{\partial\theta^2}\log f_X(x|\theta) = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}$$

Then the fisher information is

$$I(p) = -E\left[\frac{\partial^2}{\partial p^2}\log f_X(x|p)\right] = -E\left(-\frac{x}{p^2} - \frac{1-x}{(1-p)^2}\right)$$

$$= \frac{E(X)}{p^2} + \frac{1-E(X)}{(1-p)^2} = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$$

So, by Thm 10.1.12, $v(p) = p(1-p)$ and

$$\sqrt{n}(\widehat{p} - p) \xrightarrow{d} \mathcal{N}(0, p(1-p))$$

So, $\widehat{p}$ is asymptotically efficient.

6.2.3 Approximate large sample variance

We know that the asymptotic variance of $\widehat{\theta}$ is $v(\theta) = 1/I_1(\theta)$. But this is more practically the asymptotic variance of $\sqrt{n}(\widehat{\theta} - \theta)$. How do we estimate the variance of $\widehat{\theta}$ itself (without the $\sqrt{n}$ in front) in large samples using this information?

Since

$$\sqrt{n}(\widehat{\theta} - \theta) \to_d \mathcal{N}(0, v(\theta))$$

the variance of $\sqrt{n}(\widehat{\theta} - \theta)$ is close to $v(\theta)$ in large samples. So,

$$Var(\widehat{\theta}) = \frac{1}{n}Var(\sqrt{n}\widehat{\theta}) = \frac{1}{n}Var(\sqrt{n}(\widehat{\theta} - \theta)) \approx \frac{v(\theta)}{n} = \frac{1}{nI_1(\theta)} = \frac{1}{I_n(\theta)}$$

Hence, in large samples,

$$Var(\widehat{\theta}) \approx \frac{1}{I_n(\theta)}$$

Sometimes, the information as we have defined it (“expected information”) is not a very good finite sample estimate, so instead we used what is called the “observed information.”

6.2.3.1 Observed information

- When plugging in the MLE for unknown parameters in the information, we refer to this quantity as the *observed information*:

$$\left(-\frac{\partial^2}{\partial \theta^2}\log \mathcal{L}(\theta|\mathbf{x})\right)\Bigg|_{\theta=\widehat{\theta}}$$

- It has been shown that it is typically better to use observed rather than expected information for inference. (Efron & Hinkley 1978 "Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher Information")

- In general, when estimating the variance of $\tau(\widehat{\theta})$ based on an *iid* sample, we could use the observed information to approximate the variance:

$$Var(\tau(\widehat{\theta})) \approx \widehat{Var_\theta}(\tau(\widehat{\theta})) = \frac{\left(\tau'(\widehat{\theta})\right)^2}{\left(-\frac{\partial^2}{\partial\theta^2}\log\mathcal{L}(\theta|\mathbf{x})\right)\Big|_{\theta=\widehat{\theta}}}$$

- This involves *approximating* the variance with the asymptotic variance (the CRLB from Theorem 10.1.12) and then *estimating* the asymptotic variance by plugging in $\widehat{\theta}$ for the numerator as well as the observed information.
- It follows from Theorem 10.1.6 that the observed information is a consistent estimator of $I(\theta)$ so it follows that $\widehat{Var_\theta}\left[h(\widehat{\theta})\right]$ is a consistent estimator of $Var_\theta\left[h(\widehat{\theta})\right]$
- Note sometimes the expected information evaluated at the MLE, $I_n(\widehat{\theta})$, is the same as the observed information (i.e. full exponential families) at the MLE but not always.
- Using the asymptotic result for MLE's we can obtain large sample hypothesis tests and large sample confidence intervals.

In the binomial example, we can use this method of observed information to estimate the variance of functions of $\widehat{p}$, such as the MLE of the odds $\widehat{p}/(1-\widehat{p})$:

$$\begin{aligned}\widehat{Var}\left(\frac{\widehat{p}}{1-\widehat{p}}\right) &= \frac{[\frac{\partial}{\partial p}(p/(1-p))]^2|_{p=\widehat{p}}}{\left(-\frac{\partial^2}{\partial p^2}\log\mathcal{L}(p|\mathbf{x})\right)\Big|_{p=\widehat{p}}} \\ &= \frac{\left[\frac{(1-p)+p}{(1-p)^2}\right]^2|_{p=\widehat{p}}}{\frac{n}{p(1-p)}|_{p=\widehat{p}}} \\ &= \frac{\widehat{p}}{n(1-\widehat{p})^3}\end{aligned}$$

- The MLE variance approximation works well in many cases but not always.
- Since the approximation is based on the CRLB it is probably an underestimate.
- You must be careful when the function $\tau(\theta)$ is not monotone. In such cases, the derivative $\tau'(\theta)$ will have a sign change and that may lead to an underestimated variance approximation.
- Example: $\widehat{Var}(\widehat{p}(1-\widehat{p})) = \widehat{p}(1-\widehat{p})(1-2\widehat{p})^2/n$ can be 0 if $\widehat{p} = 1/2$ (need to use a second order approximation here as $p(1-p)$ is not monotone in p)
- We can also compare asymptotic variances of two estimators:

6.2.4 Asymptotic relative efficiency (ARE)

Definition 10.1.16 If two estimators W_n and V_n satisfy

$$\sqrt{n}[W_n - \tau(\theta)] \xrightarrow{d} \mathcal{N}(0, \sigma_W^2)$$

$$\sqrt{n}[V_n - \tau(\theta)] \xrightarrow{d} \mathcal{N}(0, \sigma_V^2)$$

in distribution, the *asymptotic relative efficiency (ARE)* of V_n with respect to W_n is

$$ARE(V_n, W_n) = \frac{\sigma_W^2}{\sigma_V^2}.$$

Example 10.1.17 AREs of Poisson estimators Suppose that $X_1, X_2, \ldots$ are *iid* Poisson(λ) and we are interested in estimating the 0 probability. In other words, the probability that no events happen in the time period. In this case, $P(X = 0) = e^{-\lambda}$ and a natural estimator comes from defining $Y_i = I(X_i = 0)$ and using

$$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

The Y_i's are Bernoulli($p = e^{-\lambda}$), and hence it follows from the Bernoulli distribution properties that

$$E(\widehat{\tau}) = p = e^{-\lambda}, \text{ and } Var(\widehat{\tau}) = \frac{p(1-p)}{n} = \frac{e^{-\lambda}(1-e^{-\lambda})}{n}.$$

So we can use the CLT to tell us that

$$\sqrt{n}(\widehat{\tau} - e^{-\lambda}) \to_d \mathcal{N}(0, e^{-\lambda}(1-e^{-\lambda}))$$

Alternatively, the MLE of $e^{-\lambda}$ is $e^{-\widehat{\lambda}}$ where $\widehat{\lambda} = \bar{X}_n$ is the MLE of λ. We know from the asymptotic distribution of the MLE $\widehat{\lambda}$ that:

$$\sqrt{n}(\widehat{\lambda} - \lambda) \to_d \mathcal{N}(0, v(\lambda) = 1/I_1(\lambda))$$

where

$$I_1(\lambda) = -E\left[\frac{\partial^2}{\partial\lambda^2}\log f_X(X|\lambda)\right] = -E\left[\frac{\partial^2}{\partial\lambda^2} X\log\lambda - \lambda - \log X!\right] = -E\left[\frac{\partial}{\partial\lambda}\frac{X}{\lambda} - 1\right]$$

$$= -E\left[\frac{-X}{\lambda^2}\right] = \frac{E(X)}{\lambda^2} = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}$$

so

$$\sqrt{n}(\widehat{\lambda} - \lambda) \to_d \mathcal{N}(0, \lambda)$$

Using Delta Method we have that

$$\sqrt{n}(e^{-\widehat{\lambda}} - e^{-\lambda}) \to_d \mathcal{N}(0, \lambda[-e^{-\lambda}]^2 = \lambda e^{-2\lambda})$$

Now, since

$$\sqrt{n}(\widehat{\tau} - e^{-\lambda}) \xrightarrow{d} \mathcal{N}(0, e^{-\lambda}(1-e^{-\lambda}))$$

$$\sqrt{n}(e^{-\widehat{\lambda}} - e^{-\lambda}) \xrightarrow{d} \mathcal{N}(0, \lambda e^{-2\lambda})$$

the ARE of $\widehat{\tau}$ with respect to the MLE $e^{-\widehat{\lambda}}$ is

$$ARE(\widehat{\tau}, e^{-\widehat{\lambda}}) = \frac{\lambda e^{-2\lambda}}{e^{-\lambda}(1 - e^{-\lambda})} = \frac{\lambda}{e^{\lambda} - 1}$$

This function is strictly decreasing from 1 (at $\lambda = 0$) and tails off rapidly to asymptote to 0 as $\lambda \to \infty$.

- Since the MLE is typically asymptotically efficient, another estimator cannot hope to beat its asymptotic variance.
- However, other estimators may have other desirable qualities (ease of calculation, robustness to underlying assumptions) that make them desirable.
- See C&B 10.2 for discussion on robustness
- We will cover 10.1.4 Bootstrap and 10.2 at the end of the quarter if time allows.

6.3 Hypothesis Testing (10.3)

- This section describes a few methods for deriving some tests in complicated problems.
- This is useful in settings where no optimal test (as defined in earlier sections) exists or is known (i.e. no UMP unbiased test exists).
- We will discuss large-sample properties of LRTs and other approximate large-sample tests.

6.3.1 Asymptotic Distribution of LRTs

Recall the LRT statistic is defined as:

$$\lambda(\mathbf{x}) = \frac{\sup_{\theta\in\Theta_0} \mathcal{L}(\theta|\mathbf{x})}{\sup_{\theta\in\Theta} \mathcal{L}(\theta|\mathbf{x})}$$

with rejection region $\mathcal{R} = \{\mathbf{x} : \lambda(\mathbf{x}) \leq c\}$ and a level α test we choose c such that:

$$\sup_{\theta\in\Theta_0} P_\theta(\lambda(\mathbf{X}) \leq c) \leq \alpha.$$

- Once the data $\mathbf{X} = \mathbf{x}$ are observed, the likelihood function is a completely defined function of the variable θ
- Even if the suprema cannot be analytically obtained, they can be computed numerically.
- Thus, the test statistic $\lambda(\mathbf{x})$ can be obtained for the observed data point even if there is no convenient formula for defining $\lambda(\mathbf{x})$.
- If a simple formula for $\lambda(\mathbf{x})$ cannot be derived, it will be difficult to derive the sampling distribution of $\lambda(\mathbf{X})$ in order to choose c
- However, we can use asymptotics to get an approximate answer in order to choose c

Theorem 10.3.1 Asymptotic distribution of the LRT—simple H_0 For testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ based on an *iid* sample $X_1, \ldots, X_n$ from $f(\mathbf{x}|\theta)$ satisfying the regularity conditions (Miscellanea 10.6.2). Then, under H_0 as $n \to \infty$:

$$-2\log\lambda(\mathbf{X}) \xrightarrow{d} \chi^2_1$$

(chi-square distribution with 1 degree of freedom)

Rejection of $H_0 : \theta \in \Theta_0$ for small values of $\lambda(\mathbf{X})$ is equivalent to rejection for large values of $-2\log\lambda(\mathbf{X})$:

$$\mathcal{R} = \left\{\mathbf{x} : -2\log\lambda(\mathbf{x}) \geq \chi^2_{1,\alpha}\right\}$$

Proof: Expand the log-likelihood $\log\mathcal{L}(\theta|\mathbf{x}) = \ell(\theta|\mathbf{x})$ in Taylor's series around $\widehat{\theta}$:

$$\ell(\theta|\mathbf{x}) = \ell(\widehat{\theta}|\mathbf{x}) + \ell'(\widehat{\theta}|\mathbf{x})(\theta - \widehat{\theta}) + \ell''(\widehat{\theta}|\mathbf{x})\frac{(\theta - \widehat{\theta})^2}{2} + \text{Remainder}.$$

Then substituting θ_0 for θ and multiplying by 2 gives:

$$-2\log\lambda(\mathbf{x}) = -2\ell(\theta_0|\mathbf{x}) + 2\ell(\widehat{\theta}|\mathbf{x}) = (\theta_0 - \widehat{\theta})^2(-\ell''(\widehat{\theta}|\mathbf{x})) + \text{Remainder}.$$

We can show the remainder goes to 0, and since $\ell'(\widehat{\theta}|\mathbf{x}) = 0$:

$$\begin{aligned}
-2\log\lambda(\mathbf{x}) &\approx -\ell''(\widehat{\theta})(\theta_0 - \widehat{\theta})^2 \\
&= \hat{I}(\widehat{\theta})(\theta_0 - \widehat{\theta})^2 \quad \text{observed information} \\
&= \frac{1}{n}\hat{I}(\widehat{\theta})\left[\sqrt{n}(\theta_0 - \widehat{\theta})\right]^2 \\
&\xrightarrow{d} I(\theta_0)\left[\mathcal{N}(0, 1/I(\theta_0))\right]^2 \quad \text{Slutsky's Thm \& Continuous Mapping Thm} \\
&= \left[\sqrt{I(\theta_0)}\mathcal{N}(0, 1/I(\theta_0))\right]^2 = \left[\mathcal{N}(0,1)\right]^2 = \chi^2_1
\end{aligned}$$

Hence, $-2\log\lambda(\mathbf{X}) \xrightarrow{d} \chi^2_1$.

Example 10.3.2 Poisson LRT For testing $H_0 : \lambda = \lambda_0$ versus $H_1 : \lambda \neq \lambda_0$ based on $X_1, \ldots, X_n \sim_{iid}$ Poisson(λ), we have

$$-2\log\lambda(\mathbf{x}) = -2\log\left(\frac{e^{-n\lambda_0}\lambda_0^{\sum x_i}(\prod \frac{1}{x_i!})}{e^{-n\widehat{\lambda}}\widehat{\lambda}^{\sum x_i}(\prod \frac{1}{x_i!})}\right) = 2n\left[(\lambda_0 - \widehat{\lambda}) - \widehat{\lambda}\log(\lambda_0/\widehat{\lambda})\right],$$

where $\widehat{\lambda} = \bar{x}_n$ is the MLE of λ. Applying Theorem 10.3.1, we would reject H_0 at level α if $-2\log\lambda(\mathbf{x}) > \chi^2_{1,\alpha}$.

- See table pg 490 for a simulation of how well the approximation does (pretty well) for $n = 25$
- Compare the truth = simulated percentile vs. the approximation = χ^2 percentile ($\chi^2_{1,\alpha}$)

Theorem 10.3.3 Asymptotic distribution of the LRT vector of parameters Let $X_1, \ldots, X_n$ be a random sample from $f(x|\theta)$. Under the regularity conditions (Misc 10.6.2), if $\theta \in \Theta_0$, then the distribution of the statistic $-2\log\lambda(\mathbf{X})$ converges to a chi-squared distribution as $n \to \infty$. The degrees of freedom of the limiting distribution is the difference between the number of free parameters specified by $\theta \in \Theta_0$ and the number of free parameters specified by $\theta \in \Theta$.

- Rejection of $H_0 : \theta \in \Theta_0$ for small values of $\lambda(\mathbf{X})$ is equivalent to rejection for large values of $-2\log\lambda(\mathbf{X})$:
$$\mathcal{R} = \{\mathbf{x} : -2\log\lambda(\mathbf{x}) \geq \chi^2_{\nu,\alpha}\}$$
with degree of freedom ν from above theorem.
- Type I error probability will be approximately α if $\theta \in \Theta_0$ and the sample size is large.
- An *asymptotic size α test* has the property (from above theorem):
$$\lim_{n\to\infty} P_\theta(\text{reject } H_0) = \alpha \quad \text{for each } \theta \in \Theta_0.$$
- Note this is *not* equivalent to $\lim_{n\to\infty}\sup_{\theta\in\Theta_0} P_\theta(\text{reject } H_0) = \alpha$.

Example 10.3.4 Multinomial LRT Let $\theta = (p_1, p_2, p_3, p_4, p_5)$, where $0 < p_j \leq 1$ and $\sum_{j=1}^5 p_j = 1$. Suppose $X_1, \ldots, X_n$ are *iid* discrete random variables with $P_\theta(X_i = j) = p_j, j = 1, 2, 3, 4, 5$. Thus the pmf of X_i is $f(j|\theta) = p_j$ and the likelihood function is:

$$\mathcal{L}(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta) = p_1^{y_1} p_2^{y_2} p_3^{y_3} p_4^{y_4} p_5^{y_5},$$

where y_j = number of $x_1, \dots, x_n$ equal to j, that is $y_j = \sum_{i=1}^{n} I_j(x_i)$. So the y's count the number of observations in each of the five categories.

Consider testing

$$H_0 : p_1 = p_2 = p_3 \text{ and } p_4 = p_5 \quad \text{versus} \quad H_1 : H_0 \text{ is not true.}$$

The full parameter space, Θ is really a four-dimensional set (why?) with $q = 4$ free parameters. There is only one free parameter in the set specified by H_0 since once we choose p_1 the others are determined $(p_4 = p_5 = (1 - 3p_1)/2)$. Thus Θ_0 has $p = 1$ free parameters. The degrees of freedom for the chi-square test are then $\nu = q - p = 3$.

We must calculate the LRT statistic $\lambda(\mathbf{x})$ by determining the MLE of θ under both Θ_0 and Θ (see C&B for details) to obtain the test statistic:

$$-2 \log \lambda(\mathbf{x}) = 2 \sum_{i=1}^{3} y_i \log \left(\frac{3y_i}{y_1 + y_2 + y_3} \right) + 2 \sum_{i=4}^{5} y_i \log \left(\frac{2y_i}{y_4 + y_5} \right).$$

The asymptotic size α test rejects H_0 if $-2 \log \lambda(\mathbf{x}) \geq \chi^2_{3,\alpha}$.

6.3.2 Wald Tests

We can base another large-sample test statistic on estimators with asymptotically normal distributions, such as MLEs. Remember if $X_1, \dots, X_n$ are *iid* from $f(x|\theta)$, under regularity conditions we know that the MLE of θ, $\widehat{\theta}$ satisfies:

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, v(\theta)),$$

where

$$v(\theta) = \frac{1}{I(\theta)}.$$

If $v(\theta)$ is a continuous function of θ then $v(\widehat{\theta})$ is a consistent estimator of $v(\theta)$ for all θ, that is:

$$v(\widehat{\theta}) \xrightarrow{p} v(\theta),$$

and by Slutsky's Theorem:

$$Z_n = \frac{\widehat{\theta} - \theta}{\sqrt{v(\widehat{\theta})/n}} = \frac{\widehat{\theta} - \theta}{\sqrt{v(\theta)/n}} \sqrt{\frac{v(\theta)}{v(\widehat{\theta})}} \xrightarrow{d} \mathcal{N}(0,1) \times 1 = \mathcal{N}(0,1)$$

This is the basis of the Wald test.

Definition: Wald statistic: Suppose $X_1, \ldots, X_n$ are *iid* from $f(x|\theta)$. Consider testing

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0.$$

Under H_0

$$Z_N^W = \frac{\widehat{\theta} - \theta}{\sqrt{v(\widehat{\theta})/n}} \xrightarrow{d} \mathcal{N}(0,1)$$

and so we can reject H_0 when

$$\mathcal{R} = \{\mathbf{x} : z_n^W < -z_{\alpha/2} \text{ or } z_n^W > z_{\alpha/2}\} = \{\mathbf{x} : |z_n^W| > z_{\alpha/2}\}.$$

One can also perform a one-sided test using a Wald test. For instance, if

$$H_0 : \theta \leq \theta_0 \text{ versus } H_1 : \theta > \theta_0$$

then

$$\mathcal{R} = \{\mathbf{x} : z_n^W > z_\alpha\}.$$

- More generally, instead of using the observed information $1/v(\widehat{\theta})$ we could use any consistent estimate S_n of $\sqrt{Var(\widehat{\theta})}$.

Example 10.3.5 Large-sample binomial Wald test Let $X_1, \ldots, X_n \sim_{iid}$ Bernoulli(p). Consider testing:

$$H_0 : p \leq p_0 \text{ versus } H_1 : p > p_0$$

where $0 < p_0 < 1$ is a specified value. The MLE of p is $\widehat{p} = \frac{1}{n}\sum_{i=1}^n X_i$. By the CLT and because $\widehat{p}$ is an MLE, we have

$$\sqrt{n}(\widehat{p} - p) \xrightarrow{d} \mathcal{N}(0, v(p))$$

where

$$v(p) = \frac{1}{I(p)} = p(1-p).$$

Because the asymptotic variance $v(p)$ is continuous in p we can use the observed information $v(\widehat{p})$ to consistently estimate $v(p)$ and the Wald statistic is:

$$Z_n^W = \frac{\widehat{p} - p_0}{\sqrt{\widehat{p}(1-\widehat{p})/n}}$$

The large-sample Wald test rejects H_0 if $Z_n^W > z_\alpha$:

$$\mathcal{R} = \left\{\mathbf{x} : \frac{\bar{x} - p_0}{\sqrt{\bar{x}(1-\bar{x})/n}} > z_\alpha\right\}$$

If we were interested in testing the two-sided hypothesis $H_0 : p = p_0$ versus $H_1 : p \neq p_0$, we could alternatively use $p_0(1-p_0)$ in the denominator instead of $\widehat{p}(1-\widehat{p})$. It is not clear which is preferred since the power functions cross one another and there is much discussion on this point (see pg 494).

6.3.3 Score Tests

Suppose $X_1, \dots, X_n$ are *iid* from $f(x|\theta)$. We can write the score equation function as a random variable:

$$S(\theta|\mathbf{X}) = \frac{\partial}{\partial\theta} \log \mathcal{L}(\theta|\mathbf{X}) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i|\theta) = \sum_{i=1}^{n} \ell'(\theta|X_i),$$

which is the sum of *iid* random variables. The properties of the score equations are such that:

$$E_\theta \left[\frac{\partial}{\partial\theta} \log f(X_i|\theta) \right] = 0 \tag{1}$$

$$Var_\theta \left[\frac{\partial}{\partial\theta} \log f(X_i|\theta) \right] = E_\theta \left\{ \left[\frac{\partial}{\partial\theta} \log f(X_i|\theta) \right]^2 \right\} = I(\theta). \tag{2}$$

Therefore, by the CLT:

$$\sqrt{n} \left(\frac{1}{n} S(\theta|\mathbf{X}) - 0 \right) \xrightarrow{d} \mathcal{N}(0, I(\theta))$$

and so

$$\frac{\frac{1}{n} S(\theta|\mathbf{X})}{\sqrt{I(\theta)/n}} = \frac{S(\theta|\mathbf{X})}{\sqrt{nI(\theta)}} = \frac{S(\theta|\mathbf{X})}{\sqrt{I_n(\theta)}} \xrightarrow{d} \mathcal{N}(0,1)$$

Therefore, the score function divided by the square root of Fisher's information can be approximated by a standard normal random variable. This forms the basis for the score test.

Definition: Score statistic Suppose $X_1, \dots, X_n$ are *iid* from $f(x|\theta)$ and we wish to test

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0.$$

Under H_0

$$Z_N^S = \frac{S(\theta_0|\mathbf{X})}{\sqrt{I_n(\theta_0)}} \xrightarrow{d} \mathcal{N}(0,1).$$

and so we can reject H_0 when

$$\mathcal{R} = \{\mathbf{x} : z_n^S < -z_{\alpha/2} \text{ or } z_n^S > z_{\alpha/2}\} = \{\mathbf{x} : |z_n^S| > z_{\alpha/2}\}.$$

One can also perform a one-sided test using a Score test but we need to use the MLE under the null hypothesis. For instance, if

$$H_0 : \theta \leq \theta_0 \text{ versus } H_1 : \theta > \theta_0$$

then

$$\mathcal{R} = \{\mathbf{x} : \frac{S(\widehat{\theta}_0|\mathbf{x})}{\sqrt{I_n(\widehat{\theta}_0)}} > z_\alpha\}.$$

- If H_0 is composite (i.e. a one-sided test), then $\widehat{\theta}_0$, an estimate of θ assuming H_0 is true, replaces θ_0 in Z_n^S.
- If $\widehat{\theta}_0$ is the restricted MLE, we might need to maximize using Lagrange multipliers. Thus the score test is sometimes called the Lagrange multiplier test.

Example 10.3.6 Binomial score test Consider again the test from example 10.3.5 with Bernoulli data and consider testing

$$H_0 : p = p_0 \text{ versus } H_1 : p \neq p_0.$$

The likelihood function is given by

$$\mathcal{L}(p|\mathbf{x}) = p^y(1-p)^{n-y}$$

where $y = \sum_{i=1}^n x_i$. The score equation is then:

$$S(p|\mathbf{x}) = \frac{\partial}{\partial p} \log \mathcal{L}(p|\mathbf{x}) = \frac{\partial}{\partial p} y \log p + (n-y)\log(1-p) = \frac{y}{p} - \frac{n-y}{1-p}$$

and we can use the information previously calculated (take the derivative of the score equation and the expectation):

$$I(p) = \frac{1}{p(1-p)}.$$

Therefore, the score statistic is

$$Z_n^S = \frac{S(p_0|\mathbf{X})}{\sqrt{I_n(p_0)}} = \frac{\frac{y}{p_0} - \frac{n-y}{1-p_0}}{\sqrt{\frac{n}{p_0(1-p_0)}}} = \frac{\frac{(1-p_0)y}{n} - \frac{p_0(n-y)}{n}}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\frac{y}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

where $\widehat{p} = y/n$. An approximate size α rejection region is

$$\mathcal{R} = \{\mathbf{x} : |z_n^S| > z_{\alpha/2}\}.$$

- We see that the two statistics Score vs Wald differ in only how they estimate the standard error of $\widehat{p}$: The Wald statistic uses the estimated standard error. The score statistic uses the standard error calculated under the assumption that $H_0 : p = p_0$ is true.

6.3.4 Summary of asymptotic test statistics

If we have $X_1, \ldots, X_n \sim_{iid} f(x|\theta)$ and we assume the regularity conditions needed for MLEs to be consistent and asymptotically normal hold. We have three large sample procedures to test the hypothesis:

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0.$$

- **LRT:**
$$-2 \log \lambda(\mathbf{X}) = -2[\log \mathcal{L}(\theta_0|\mathbf{X}) - \log \mathcal{L}(\widehat{\theta}|\mathbf{X})] \xrightarrow{d} \chi^2_1$$
- **Wald:**
$$Z_n^W = \frac{\widehat{\theta} - \theta_0}{\sqrt{\frac{v(\widehat{\theta})}{n}}} = \frac{\widehat{\theta} - \theta_0}{\sqrt{\frac{1}{I_n(\widehat{\theta})}}} \xrightarrow{d} \mathcal{N}(0, 1)$$
- **Score:**
$$Z_n^S = \frac{S(\theta_0|\mathbf{X})}{\sqrt{I_n(\theta_0)}} \xrightarrow{d} \mathcal{N}(0, 1)$$

All convergence results are under $H_0 : \theta = \theta_0$.

- Note that $(Z_n^W)^2, (Z_n^S)^2$, and $-2 \log \lambda(\mathbf{X})$ each converge in distribution (under the null hypothesis) to a χ^2_1 distribution as $n \to \infty$.

6.4 Interval Estimation (10.4)

- In Chapter 9, we discussed methods to derive confidence intervals based on exact (i.e. finite sample) distributions.
- In complicated situations, we may need to resort to asymptotic theory to develop large sample approximate confidence intervals using the large-sample approximate estimators and tests above (Wald, Score, LRT).
- These are known as the “large sample likelihood based confidence intervals."
- We can invert asymptotic tests of size α to obtain confidence intervals with asymptotic coverage $1 - \alpha$.
- We can also use asymptotic (large sample) pivots.

Definition Suppose $X_1, \ldots, X_n \sim_{iid} f(x|\theta)$. The random variable

$$Q_n = Q_n(\mathbf{X}, \theta)$$

is called a *large sample pivot* if its asymptotic distribution is free of all unknown parameters. If Q_n is a large sample pivot and if

$$P_\theta(Q_n(\mathbf{X}, \theta) \in \mathcal{A}) \approx 1 - \alpha,$$

then

$$C(X) = \{\theta : Q_n(\mathbf{X}, \theta) \in \mathcal{A}\}$$

is called an *approximate* $1 - \alpha$ *confidence set* for θ.

6.4.1 Wald intervals - pivot

For data $X_1, \ldots, X_n \sim_{iid} f(x|\theta)$ under regularity conditions we know that the MLE $\widehat{\theta}$ is asymptotically normal and efficient:

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, v(\theta)),$$

where

$$v(\theta) = \frac{1}{I_1(\theta)}.$$

If $v(\theta)$ is a continuous function of θ, then $v(\widehat{\theta}) \xrightarrow{p} v(\theta)$ for all θ ($v(\widehat{\theta})$ is a consistent estimator of $v(\theta)$) and

$$Q_n(\mathbf{X}, \theta) = \frac{\widehat{\theta} - \theta}{\sqrt{\frac{v(\widehat{\theta})}{n}}} \xrightarrow{d} \mathcal{N}(0, 1),$$

by Slutsky's Theorem. Therefore, $Q_n(\mathbf{X}, \theta)$ is a large sample pivot and

$$\begin{aligned} 1 - \alpha &\approx P_\theta(-z_{\alpha/2} \le Q_n(\mathbf{X}, \theta) \le z_{\alpha/2}) \\ &= P_\theta(-z_{\alpha/2} \le \frac{\widehat{\theta} - \theta}{\sqrt{\frac{v(\widehat{\theta})}{n}}} \le z_{\alpha/2}) \\ &= P_\theta\left(\widehat{\theta} - z_{\alpha/2}\sqrt{\frac{v(\widehat{\theta})}{n}} \le \theta \le \widehat{\theta} + z_{\alpha/2}\sqrt{\frac{v(\widehat{\theta})}{n}}\right). \end{aligned}$$

Therefore,

$$\widehat{\theta} \pm z_{\alpha/2}\sqrt{\frac{v(\widehat{\theta})}{n}}$$

is an approximate $1 - \alpha$ confidence interval for θ.

6.4.2 Wald intervals - inversion of hypothesis test

We could have approached this problem from a different direction by inverting the large sample test of

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \ne \theta_0$$

that uses the Wald test statistic

$$Z_n^W = \frac{\widehat{\theta} - \theta_0}{\sqrt{\frac{v(\widehat{\theta})}{n}}}$$

and rejection region

$$\mathcal{R} = \{x : |z_n^W| \ge z_{\alpha/2}\}$$

So we can invert the test to obtain the CI:

$$\begin{aligned} CI &= \{\theta : |z_n^W| \le z_{\alpha/2}\} \\ &= \{-z_{\alpha/2} \le \frac{\widehat{\theta} - \theta}{\sqrt{v(\widehat{\theta})/n}} \le z_{\alpha/2}\} \\ &= \{-z_{\alpha/2}\sqrt{v(\widehat{\theta})/n} \le \widehat{\theta} - \theta \le z_{\alpha/2}\sqrt{v(\widehat{\theta})/n}\} \\ &= \{\widehat{\theta} - z_{\alpha/2}\sqrt{v(\widehat{\theta})/n} \le \theta \le \widehat{\theta} + z_{\alpha/2}\sqrt{v(\widehat{\theta})/n}\} \end{aligned}$$

This is why this type of large sample interval is called a *Wald confidence interval* as it is the interval that arises from inverting a large sample Wald test.

6.4.3 Wald intervals - delta method

We can also create large sample Wald confidence intervals for functions of θ using the Delta Method. From the delta method if we have $g(\theta)$ such that g' exists and is nonzero, then

$$\sqrt{n}(g(\widehat{\theta}) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, [g'(\theta)]^2 v(\theta)).$$

If $[g'(\theta)]^2 v(\theta)$ is a continuous function of θ, then $[g'(\widehat{\theta})]^2 v(\widehat{\theta})$ is a consistent estimator for it (continuous function of consistent MLEs also consistent). Therefore,

$$Q_n(\mathbf{X}, \theta) = \frac{g(\widehat{\theta}) - g(\theta)}{\sqrt{\frac{[g'(\widehat{\theta})]^2 v(\widehat{\theta})}{n}}} \xrightarrow{d} \mathcal{N}(0, 1).$$

by Slutsky's Theorem and

$$g(\widehat{\theta}) \pm z_{\alpha/2}\sqrt{\frac{[g'(\widehat{\theta})]^2 v(\widehat{\theta})}{n}}$$

is an approximate $1 - \alpha$ confidence interval for $g(\theta)$.

Example Bernoulli Wald interval Suppose $X_1, \ldots, X_n \sim_{iid} Bernoulli(p)$. To derive a $1 - \alpha$ large sample Wald confidence interval for p, we need the MLE of p: $\widehat{p} = \frac{1}{n}\sum_{i=1}^n X_i$. Earlier we showed that

$$v(p) = \frac{1}{I(p)} = p(1 - p).$$

Therefore,

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1 - \widehat{p})}{n}}$$

is an approximate $1 - \alpha$ Wald confidence interval for p. This interval has problems (i.e. inability to attain the nominal $1 - \alpha$ coverage probability), see Brown et al. (2001, *Statistical Science*).

Now, if we wish to obtain a $1-\alpha$ large sample Wald confidence interval for the log odds of p

$$g(p) = \log\left(\frac{p}{1-p}\right)$$

we can use the Delta method since

$$g'(p) = \frac{1}{p(1-p)} \neq 0 \text{ for } 0 < p < 1$$

Therefore,

$$\begin{aligned}\sqrt{n}\left[\log\left(\frac{\widehat{p}}{1-\widehat{p}}\right) - \log\left(\frac{p}{1-p}\right)\right] &\xrightarrow{d} \mathcal{N}\left(0, \left[\frac{1}{p(1-p)}\right]^2 p(1-p)\right)\\ &= \mathcal{N}\left(0, \frac{1}{p(1-p)}\right).\end{aligned}$$

Because the asymptotic variance $1/(p(1-p))$ can be consistently estimated by $1/(\widehat{p}(1-\widehat{p}))$, we have

$$\frac{\log\left(\frac{\widehat{p}}{1-\widehat{p}}\right) - \log\left(\frac{p}{1-p}\right)}{\sqrt{\frac{1}{n\widehat{p}(1-\widehat{p})}}} \xrightarrow{d} \mathcal{N}(0,1)$$

by Slutsky's Theorem, and

$$\log\left(\frac{\widehat{p}}{1-\widehat{p}}\right) \pm z_{\alpha/2}\sqrt{\frac{1}{n\widehat{p}(1-\widehat{p})}}$$

is an approximate $1-\alpha$ Wald confidence interval for $g(p) = \log[p/(1-p)]$.

- Clearly, the Wald interval is simple and straightforward. All we need is an MLE and a consistent estimator of the asymptotic variance of the MLE.
- More generally, to perform Wald inference all you need is an estimator $\widehat{\theta}$ (not necessarily an MLE) that is asymptotically normal with a large sample variance that you can estimate consistently.
- However, because large sample standard errors must be estimated, the performance of Wald confidence intervals and tests can be poor in small samples.
- Wald inference is really a last resort unless you have very large samples (i.e. thousands of subjects), in which case it is often used because of its simplicity.

6.4.4 Score intervals

For data $X_1, \ldots, X_n \sim_{iid} f(x|\theta)$ under regularity conditions we have shown that

$$Q_n(\mathbf{X}, \theta) = \frac{S(\theta|\mathbf{X})}{\sqrt{I_n(\theta)}} \xrightarrow{d} \mathcal{N}(0,1)$$

where $I_n(\theta) = nI(\theta)$ is the Fisher information based on the sample.

Score confidence intervals arise from inverting (large sample) score tests. When testing

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0$$

the score statistic

$$Q_n(\mathbf{X}, \theta_0) = \frac{S(\theta_0|\mathbf{X})}{\sqrt{I_n(\theta_0)}} \xrightarrow{d} \mathcal{N}(0,1)$$

when H_0 is true. Therefore,

$$\mathcal{R} = \{\mathbf{x} : |Q_n(\mathbf{x}, \theta_0)| \geq z_{\alpha/2}\}$$

is an approximate size α rejection region for testing H_0 versus H_1

We can invert the score test by finding the acceptance region

$$\mathcal{A} = \{\mathbf{x} : |Q_n(\mathbf{x}, \theta_0)| < z_{\alpha/2}\}$$

and inverting it to obtain the approximate $1 - \alpha$ confidence set for θ:

$$C(\mathbf{x}) = \{\theta : |Q_n(\mathbf{x}, \theta)| < z_{\alpha/2}\}$$

When $C(\mathbf{x})$ is an interval, this is the *score confidence interval.*

Example Bernoulli Score interval Again we have $X_1, \ldots, X_n \sim_{iid} Bernoulli(p)$ where $0 < p < 1$. We have the score test statistic:

$$Q_n(\mathbf{X}, p) = \frac{S(p|\mathbf{X})}{\sqrt{I_n(p_0)}} = \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

The score test of $H_0 : p = p_0$ vs $H_1 : p \neq p_0$ uses this test statistic with $p = p_0$. We invert the acceptance region of this test as above to create the random confidence set:

$$C(\mathbf{X}) = \{p : Q_n(\mathbf{X}, p) < z_{\alpha/2}\} = \left\{ p : \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2} \right\}$$

This is the score interval for p.

- Note that it is not solved for p analytically.
- After observing $\mathbf{X} = \mathbf{x}$ we can calculate this interval numerically (using a grid search for p that satisfy this inequality).
- However, in the binomial case, we *can* get a closed-form expression for the endpoints by solving $Q_n(\mathbf{x}, p) = z_{\alpha/2}$ with the quadratic formula. See the long expression for the endpoints (10.4.7) in C&B.

6.4.5 Likelihood ratio intervals

We can also invert the likelihood ratio test for the hypothesis

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0$$

with the LRT statistic

$$\lambda(\mathbf{x}) = \frac{\mathcal{L}(\theta_0|\mathbf{x})}{\mathcal{L}(\widehat{\theta}|\mathbf{x})}$$

and rejection region

$$\mathcal{R} = \{\mathbf{x} : -2\log\lambda(\mathbf{x}) \geq \chi^2_{1,\alpha}\}.$$

$$\mathcal{A} = \{\mathbf{x} : -2\log\frac{\mathcal{L}(\theta_0|\mathbf{x})}{\mathcal{L}(\widehat{\theta}|\mathbf{x})} < \chi^2_{1,\alpha}\}.$$

We obtain an approximate size α rejection region and inverting the acceptance region gives the confidence set:

$$C(\mathbf{x}) = \{\theta : -2\log\left[\frac{\mathcal{L}(\theta|\mathbf{x})}{\mathcal{L}(\widehat{\theta}|\mathbf{x})}\right] < \chi^2_{1,\alpha}\}$$

which is an approximate $1-\alpha$ confidence set, and if it is an interval it is the *likelihood ratio confidence interval.*

Example Bernoulli LRT interval Under the same settings as above, we have the LRT statistic:

$$\lambda(\mathbf{x}) = -2\log\frac{\mathcal{L}(p_0|\mathbf{x})}{\mathcal{L}(\widehat{p}|\mathbf{x})} = -2\left[n\widehat{p}\log\frac{p_0}{\widehat{p}} + n(1-\widehat{p})\log\frac{1-p_0}{1-\widehat{p}}\right]$$

so the confidence interval is

$$C(\mathbf{x}) = \left\{p : -2\left[n\widehat{p}\log\frac{p}{\widehat{p}} + n(1-\widehat{p})\log\frac{1-p}{1-\widehat{p}}\right] < \chi^2_{1,\alpha}\right\}$$

which must be calculated using numerical search methods.

6.4.6 Comparison of binomial intervals

For $n = 12$, C&B compares 90% Score, Wald, and LRT confidence intervals for the bernoulli proportion.

- the Score interval is longer
- the Score interval is the only interval that maintains coverage above nominal level (confidence coefficient = 0.9)
- near the boundaries p close to 0 or 1, the LRT coverage drops toward 0.7
- the Wald procedure performs poorly (coverage near 0.7, centered at $\widehat{p}$, usually longer unless near boundary) for many values of p
- See Figure 10.4.2
- continuity corrected Score interval performs the best in this situation (see C&B pg 105 for definition)