

Honours Statistics
Professor Wolfson
McGill University
Math 357

Daphna Harel

Winter 2009

Contents

1	Probability: Odds and Ends	5
1.1	Introduction	5
1.1.1	Examples of the use of Statistics	5
1.2	Multivariate transformations	6
1.2.1	Transformation Theorems	6
1.2.2	Order Statistics	9
1.2.3	Multinomial Distribution	11
1.2.4	Application to Order Statistics	13
1.2.5	The Sampling Distributions	13
2	Statistical Inference	20
2.1	Introduction	20
2.2	Properties of Estimators	22
2.2.1	Unbiasedness and Minimum Variance	22
2.2.2	Consistency	28
2.2.3	Sufficiency	30
2.2.4	Completeness and a Return to Minimum Variance Unbiasedness	37
2.3	Methods of Estimation	40
2.3.1	Method of Moments	40
2.3.2	Method of Maximum Likelihood	42
2.3.3	Invariance and MLEs	45
2.4	Interval Estimation	48
2.4.1	Frequentist Interpretations	49
2.4.2	The Actual Computation	50
2.4.3	The Two-Sample Setup	50
2.4.4	Sample Size Calculation	53
2.4.5	Opinion Polls:	54
2.5	Hypothesis (or Significance) Testing	55
2.5.1	The Construction of Optimal Tests Under Certain Circumstances	59
2.5.2	The Likelihood Ratio Test	64
2.5.3	p-Values	66

3	The Bayesian Approach to Statistics	70
3.1	Similarities and Differences	70
3.2	Hypothesis Testing	72

1 Probability: Odds and Ends

1.1 Introduction

Much of statistics is concerned with making inference (i.e. making an educated guess) about a population based on a sample drawn from the population. The population will be represented by a distribution, F , describing the values in the population.

F may be parametrically specified, for example, $F \sim N(\mu, \sigma^2)$, where μ and σ^2 are unknown. A statistical problem might then be, “Make inference about μ and σ^2 , based on a sample $X_1, X_2, \dots, X_n$ drawn from F .” In general, “make inference about θ drawn from $X = (X_1, \dots, X_n)$ drawn from $F(x; \theta)$ where θ is unknown. Such problems are called **parametric**, since what is unknown is a finite set of parameter values.

On the other hand, if we are not prepared to specify a parametric family, then we are left with a problem of inference about an unknown F , where F could be any distribution. Such problems are termed **non-parametric** and will mostly be excluded from these notes.

Since we shall make inference about θ based on a random sample $(X_1, \dots, X_n)$ drawn from $F(x; \theta)$, we can never be certain about our “guess” at θ . Hence, probability is used to quantify the possible error in making a guess at θ .

1.1.1 Examples of the use of Statistics

1. Opinion Polls
2. Clinical Trials
3. Optimal Design
4. Epidemiological Data
5. Vaccine Assessment
6. DNA fingerprinting

1.2 Multivariate transformations

The following is an extension of the ideas for univariate transformations:

1.2.1 Transformation Theorems

Theorem 1.1. *Let $X = (X_1, \dots, X_n) \sim f_X(x)$, where f_X is a p.d.f. that is continuous except possibly at a finite number of points of S , the support of f_X . Let $Y = h(X)$ be a transformation such that h is one-to-one on S and onto $T = h(S)$ (i.e. h^{-1} exists and is unique). Let $h^{-1} = g$.*

Thus we have:

$$\begin{array}{ll} Y_1 = h_1(x_1, \dots, x_n) & X_1 = g_1(y_1, \dots, y_n) \\ Y_2 = h_2(x_1, \dots, x_n) & X_2 = g_2(y_1, \dots, y_n) \\ \vdots & \vdots \\ Y_n = h_n(x_1, \dots, x_n) & X_n = g_n(y_1, \dots, y_n) \end{array}$$

Suppose that $\frac{\partial g_i}{\partial y_j}(y_1, \dots, y_n)$ exists and is continuous $\forall y \in T, i, j = 1, 2, \dots, n$.

Let $J = \det \left\{ \frac{\partial g_i}{\partial y_j}(y_1, \dots, y_n) \right\}$ (The so-called jacobian).

Then,

$$f_Y(y) = \begin{cases} f_X(g_1(y_1, \dots, y_n), g_2(y_1, \dots, y_n), \dots, g_n(y_1, \dots, y_n)) \cdot |J| & \text{for } y \in T \\ 0 & \text{else} \end{cases}$$

Application: Let X_1, X_2 be i.i.d $N(0,1)$ random variables. Let $Y_1 = \frac{X_1+X_2}{2}, Y_2 = \frac{X_1-X_2}{2}$.

- i. Show that Y_1 and Y_2 are independent
- ii. Find the distribution of Y_1 and Y_2

Solution:

Intuitively, we express the X_i 's in terms of the Y_i 's, then substitute these expressions into the PDF and multiply by the Jacobian as a "correction."

Here we have, for $-\infty \leq X_1, X_2 \leq \infty$:

$$f_X(x) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}X_1^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}X_2^2}$$

And therefore:

$$S = \mathbb{R} \times \mathbb{R}$$

$$T = h(S) = \mathbb{R} \times \mathbb{R}$$

$$Y_1 = \frac{X_1 + X_2}{2} = h_1(x_1, x_2)$$

$$Y_2 = \frac{X_1 - X_2}{2} = h_2(x_1, x_2)$$

So, we set $g_1(Y_1, Y_2) \equiv X_1 = Y_1 + Y_2$ and $g_2(Y_1, Y_2) \equiv X_2 = Y_1 - Y_2$.

All the first and second partial derivatives exist and are continuous, so we calculate the Jacobian matrix, J , and take its determinant.

$$\text{In this case, } |\det(J)| = \left| \det \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \right| = |-2| = 2$$

Finally, use the transformation theorem to get, for $y \in \mathbb{R} \times \mathbb{R}$:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_1+y_2)^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_1-y_2)^2} \cdot 2 = \frac{1}{\pi} e^{-y_1^2} e^{-y_2^2}$$

So if we set:

$$\begin{aligned} f_1(y_1) &= \sqrt{2} \frac{1}{\sqrt{2\pi}} e^{-y_1^2} & -\infty < y_1 < \infty \\ f_2(y_2) &= \sqrt{2} \frac{1}{\sqrt{2\pi}} e^{-y_2^2} & -\infty < y_2 < \infty \end{aligned}$$

Then it is clear that f_1 and f_2 are $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$ respectively and therefore Y_1 and Y_2 are independent. In fact, we recognize that $Y_1 \sim N(0, \frac{1}{2})$ and $Y_2 \sim N(0, \frac{1}{2})$.

Alternatively, we could see that Y_1 is normal since $\frac{X_1}{2}$ and $\frac{X_2}{2}$ are independent normal random variables, and therefore their sum must be normal. Furthermore, $\mathbb{E}(\frac{X_1+X_2}{2}) = 0$ and $\text{Var}(\frac{X_1+X_2}{2}) = \frac{1}{4}[\text{Var}(X_1)+\text{Var}(X_2)] = \frac{2}{4} = \frac{1}{2}$

Comments:

1. Naturally, $X_1 + X_2$ and $X_1 - X_2$ are also independent (that is, functions of independent random variables are independent)
2. It is clear that if X_1, X_2 are independent $N(\mu, \sigma^2)$, in general, Y_1 and Y_2 as defined above will still be independent (Set $\frac{X_1-\mu}{\sigma} = W_1 \sim N(0,1)$ and $\frac{X_2-\mu}{\sigma} = W_2 \sim N(0,1)$, and we know from above that $W_1 \perp W_2$)

From above,

$$\begin{aligned} \frac{W_1 + W_2}{2} \perp \frac{W_1 - W_2}{2} &\iff \frac{X_1 - \mu}{2\sigma} + \frac{X_2 - \mu}{2\sigma} \perp \frac{X_1 - \mu}{2\sigma} - \frac{X_2 - \mu}{2\sigma} \\ &\iff \frac{1}{2\sigma}[(X_1 + X_2) - 2\mu] \perp \frac{1}{2\sigma}[(X_1 - X_2)] \end{aligned}$$

Therefore, $\frac{X_1+X_2}{2}$ and $\frac{X_1-X_2}{2}$ are independent since they are functions of independent random variables.

Generalization of the Transformation Theorem:

Theorem 1.2. *Let $X \sim f_X(x)$, where the p.d.f f_X is continuous except possibly at a finite number of points of S , the support of f_X . Let $\underline{Y} = (Y_1, \dots, Y_n) = h(\underline{X}) = h(X_1, \dots, X_n)$. Suppose that:*

1. *h is not necessarily 1 - 1 from S onto $T = h(S)$*
2. *We can write $S = \bigcup_{j=1}^k S_j$, where $S_j \cap S_l = \emptyset$ for $j \neq l$ (We partition the support into finitely many sections on which the function is one-to-one)*
3. *The restriction of h to S_j is called h_j (note that here, j refers to the partition index and NOT to a vector component), where h_j is assumed to be 1 - 1 from S_j to $T_j = h_j(S_j)$.*
4. *$\frac{\partial g_{ij}}{\partial y_l}(y_1, \dots, y_n)$ exists and is continuous for $i, l = 1, \dots, n$ and each $j = 1, \dots, k$, where $g_j \equiv h_j^{-1}$*

Define $J_j = \det \left\{ \frac{\partial g_{ij}}{\partial y_l}(y_1, \dots, y_n) \right\}$.

Then,

$$f_Y(y) = \begin{cases} \sum_{j=1}^k f_X(g_{j1}(y_1, \dots, y_n), \dots, g_{jn}(y_1, \dots, y_n)) \cdot |J_j| & \text{for } y \in T \\ 0 & \text{else} \end{cases}$$

Note: This result is analogous to the single-variable case, where h is not one-to-one. We apply the previous proposition to each partition, and then sum over the partitions.

1.2.2 Order Statistics

The following is an example of the transformation theorem:

Let $X_1, X_2, \dots, X_n$ be jointly continuous random variables with joint density $f_X(x)$.

Let:

$$\begin{aligned} Y_1 &= \min(X_1, \dots, X_n) \\ Y_2 &= \min(\{X_1, \dots, X_n\} \setminus \{Y_1\}) \\ &\vdots \\ Y_n &= \max(X_1, \dots, X_n) \end{aligned}$$

With probability 1, we have $Y_1 < Y_2 < \dots < Y_n$.

Definition: $Y_1, Y_2, \dots, Y_n$ are called the n order statistics of $X_1, X_2, \dots, X_n$.

(Alternative notation: $X_{(1)} < X_{(2)} < \dots < X_{(n)}$)

Y_i is called the i^{th} order statistic, Y_1 is the minimum and Y_n is the maximum.

As we will see, the order statistics play an important role in the theory and practice of statistics. The joint density $f_Y(\underline{y})$ of the order statistics $Y_1 < Y_2 < \dots < Y_n$ is of fundamental importance.

Theorem 1.3. Let $X_1, X_2, \dots, X_n$ be jointly continuous, i.i.d random variables, each with p.d.f. f_{X_1} , such that f_X has at most a finite number of discontinuities.

Then,

$$f_Y(y_1, y_2, \dots, y_n) = \begin{cases} n! \prod_{i=1}^n f_{X_1}(y_i) & \text{if } y_1 < y_2 < \dots < y_n \\ 0 & \text{otherwise} \end{cases}$$

Proof. First, observe that $\forall x \in S$ (i.e. $\forall x_i \in S_i$ of f_{X_1} , since we have the same S_i 's):

$$f_X(x_1, \dots, x_n) = \prod_{i=1}^n f_{X_1}(x_i),$$

Define $h(\underline{x}) = \underline{y}$ as follows:

For each $y_1 < y_2 < \dots < y_n$ and for each $(x_1, \dots, x_n)$

Set:

$$y_1 = \min(x_1, \dots, x_n)$$

$$\vdots$$

$$y_n = \max(x_1, \dots, x_n)$$

Consider the set of all permutations of $(x_1, \dots, x_n)$, of which there are $n!$. We see that the transformation $h(x) = y$ is not injective, since any permutation of the $(x_1, \dots, x_n)$ gives the same $(y_1, \dots, y_n)$. However, for each fixed permutation, we do have a one-to-one mapping onto $(y_1, \dots, y_n)$.

We apply the generalization of the transformation theorem:

Assume that f_{X_1} is continuous except at possibly a finite number of points.

$$S_j = (x_1, x_2, \dots, x_n)_j = j^{th} \text{ permutation of } (x_1, x_2, \dots, x_n).$$

$$T_j = (y_1, y_2, \dots, y_n) \quad \forall j = 1, \dots, n!$$

It is clear that $\frac{\partial g_{ij}}{\partial y_l}(y_1, \dots, y_n)$ exist and are continuous for each $j, i, l = 1, \dots, n$ and that all partials are either 0 or 1. Therefore, $|J_j| = 1$.

Hence, for $y_1 < y_2 < \dots < y_n$

$$f_Y(y) = \sum_j \prod_{i=1}^n f_{X_1}(y_i) = n! \prod_{i=1}^n f_X(y_i)$$

□

With the joint p.d.f. of the Y 's, one can derive the joint p.d.f. of any subset of the Y 's.

Example: Let $X_1, X_2, \dots, X_5$ have joint p.d.f. $\prod_{i=1}^5 f_X(x_i)$. Given $f_{Y_1, \dots, Y_5}(y_1, \dots, y_5)$, find $f_{Y_1, Y_2, Y_4}(y_1, y_2, y_4)$.

Solution:

$$\begin{aligned} f_{Y_1, Y_2, Y_4}(y_1, y_2, y_4) &= 5! \int_{y_2}^{y_4} \int_{y_4}^{\infty} f_{X_1}(y_1) f_{X_1}(y_2) f_{X_1}(y_3) f_{X_1}(y_4) f_{X_1}(y_5) dy_5 dy_3 \\ &= 5! \cdot f_{X_1}(y_1) f_{X_1}(y_2) f_{X_1}(y_4) \int_{y_2}^{y_4} f_{X_1}(y_3) (1 - F_{X_1}(y_4)) dy_3 \\ &= 5! \cdot f_{X_1}(y_1) f_{X_1}(y_2) f_{X_1}(y_4) (F_{X_1}(y_4) - F_{X_1}(y_2)) (1 - F_{X_1}(y_4)) \end{aligned}$$

1.2.3 Multinomial Distribution

The multinomial distribution is a generalization of the binomial distribution. As we will see, the multinomial distribution will provide us with a heuristic way of deriving the joint marginals that requires no integration.

Definition: $X_1, \dots, X_n$ have a *multinomial distribution* with parameters, n and $p_1, p_2, \dots, p_k$

with $\sum_{i=1}^k p_i = 1$ where $p_i \geq 0$, if:

$$P_{(X_1, \dots, X_k)}(x_1, \dots, x_k) = \Pr(X_1 = x_1, \dots, X_k = x_k)$$

$$= \begin{cases} \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}, & \text{when } \sum_{i=1}^k x_i = n \\ 0 & \text{otherwise} \end{cases}$$

It is obvious that the multinomial distribution reduces to the binomial when $k = 2$. Also, observe that $p_k = 1 - \sum_{i=1}^{k-1} p_i$ and that $x_k = n - \sum_{i=1}^{k-1} x_i$.

As with the Binomial, the multinomial most often arises in a certain setup:

Multinomial Setup:

1. Suppose that we have n independent trials
2. Suppose that each trial can result in exactly one of k possible outcomes
3. Suppose that the probability of outcome i is p_i , for $i = 1, \dots, k$ independent of the trial number.

Let $X_1, X_2, \dots, X_k$ represent the numbers of Type 1, Type 2, ..., Type k outcomes respectively in the n trials, then:

Claim: For $\sum_{i=1}^n p_i = 1$, and for $\sum_{i=1}^k x_i = n$,

$$P(X_1 = x_1, \dots, X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} \cdot p_1^{x_1} \cdots p_k^{x_k}$$

Proof. We simply follow the same steps as for the binomial, arguing from a specific configuration to the union of all configurations which lead to $x_1, x_2, \dots, x_k$ of Type 1, ..., Type k . □

1.2.4 Application to Order Statistics

Example: Let $X_1, X_2, \dots, X_{10}$. Find the joint densities of (Y_2, Y_4, Y_8, Y_{10}) .

We will find:

$$P(Y_2 \in (y_2, y_2 + \Delta y_2], Y_4 \in (y_4, y_4 + \Delta y_4], Y_8 \in (y_8, y_8 + \Delta y_8], Y_{10} \in (y_{10}, y_{10} + \Delta y_{10}]) \\ (\star) \approx f_{Y_2, Y_4, Y_8, Y_{10}}(y_2, y_4, y_8, y_{10}) dy_2 dy_4 dy_8 dy_{10}$$

The idea is to create a multinomial setup with “boxes,” where the first box is 0 to y_2 , the second box at y_2 , then a box from y_2 until y_4 , and so on.

Obviously, for $y_2 < y_4 < y_8 < y_{10}$,

$$(\star) = P(1 \in B_1, 1 \in B_2, 1 \in B_3, 1 \in B_4, 3 \in B_5, 1 \in B_6, 1 \in B_7, 1 \in B_8, 0 \in B_9) \\ = \frac{10!}{1!1!1!1!3!1!1!1!0!} \cdot (F_{X_1}(y_2))^1 \cdot (f_{X_1}(y_2)dy_2)^1 \cdot (F_{X_1}(y_4) - F_{X_1}(y_2))^1 \cdot (f_{X_1}(y_4)dy_4)^1 \\ \cdot (F_{X_1}(y_8) - F_{X_1}(y_4))^3 \cdot (f_{X_1}(y_8)dy_8)^1 \cdot (F_{X_1}(y_{10}) - F_{X_1}(y_8))^1 \cdot (f_{X_1}(y_{10})dy_{10})^1 \cdot (1 - F_{X_1}(y_{10}))^0$$

And if we use the “i.i.d.-ness,” we see the multinomial setup for the trials involving the X ’s.

1.2.5 The Sampling Distributions

The following section is fundamental as it concerns the distribution of functions of the sample of observations, $(X_1, X_2, \dots, X_n)$

Theorem 1.4. Let $X_1, X_2, \dots, X_n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$. Then

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1), \text{ where } \bar{X} \equiv \frac{1}{n} \sum_{i=1}^n X_i$$

Proof. The m.g.f. of a $N(\mu, \sigma^2)$ random variable is $M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$. We now use that for i.i.d. random variables, $M_{X+Y}(t) = M_X(t) \cdot M_Y(t)$ and that $M_{aX+b}(t) = e^{bt} M_X(at)$, and therefore, we get the m.g.f. for a $N(0, 1)$ distribution. $\square$

Theorem 1.5. (Central Limit Theorem) Let $X_1, X_2, \dots, X_n$ be independent, identically distributed random variables with mean μ and variance σ^2 . Then,

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

Theorem 1.6. If $X_n \xrightarrow{\mathcal{D}} X$ and $Y_n \xrightarrow{P, \mathcal{D}} c$ then $Y_n X_n \xrightarrow{\mathcal{D}} cX$.

Lemma 1.6.5

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

Proof.

$$\begin{aligned} \sum_{i=1}^n (X_i - \bar{X})^2 &= \sum_{i=1}^n \left[X_i^2 - 2X_i \sum X_i \cdot \frac{1}{n} + \left(\frac{1}{n} \sum X_i \right)^2 \right] \\ &= \sum X_i^2 - \frac{2}{n} \sum X_i \sum X_i + n\bar{X}^2 \\ &= \sum X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 \\ &= \sum_{i=1}^n X_i^2 - n\bar{X}^2 \end{aligned}$$

□

Theorem 1.7. Let $X_1, X_2, \dots, X_n$ be i.i.d. random variables with finite mean $\mu < \infty$ and finite variance $\sigma^2 < \infty$.

Define:

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

Then

$$s^2 \xrightarrow{P} \sigma^2$$

Proof. Warning: Do not try to use Markov's Inequality, since you would need to assume the existence of 4 moments.

It is much easier if we note that as $n \rightarrow \infty$:

$$\begin{aligned} \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 &= \frac{1}{n-1} \left[\sum_{i=1}^n (X_i^2) - n\bar{X}^2 \right] \\ &= \frac{1}{n-1} \sum_{i=1}^n X_i^2 - \frac{n}{n-1} \bar{X}^2 \\ &\xrightarrow{P} \mathbb{E}(X_i^2) - \mu^2 \text{ (by WLLN)} \\ &= \sigma^2 \end{aligned}$$

□

Theorem 1.8. *If $X_1, X_2, \dots, X_n$ are i.i.d. with finite μ and σ , then:*

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \xrightarrow{\mathcal{D}} Z \sim N(0, 1) \text{ as } n \rightarrow \infty$$

Proof.

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}} \cdot \frac{s}{\sigma}} \xrightarrow{\mathcal{D}} Z \cdot 1$$

(since by Theorem 1.7 $\frac{s}{\sigma} \rightarrow 1$)

□

Comment: The convergence is uniform in x .

That is:

$$P(Y_n \leq x) \longrightarrow P(Z \leq x) \text{ uniformly in } x.$$

Theorem 1.9. *Let $X_1, X_2, \dots, X_n$ be i.i.d. $N(\mu, \sigma^2)$*

Then:

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

*i.e. the left hand side has a **student t-distribution** with **$n-1$ degrees of freedom***

Aside About the t-Distribution:

1. The exact form of the p.d.f. is not important
2. It is symmetric about 0. It looks like a $N(0,1)$ but has heavier tails [that is, the tails do not approach 0 as quickly as an $N(0,1)$]

3. As $n \rightarrow \infty$, the t-distribution approaches a $N(0,1)$ distribution

We prove the following three theorems in preparation for the proof of Theorem 1.9.

Theorem 1.10. *Let $U \sim N(0, 1)$, $V \sim \chi_r^2$. Let U and V be independent.*

Then $\frac{U}{\sqrt{\frac{V}{r}}} \sim t_r$

Theorem 1.11.

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$$

Proof.

$$\underbrace{\frac{1}{\sigma^2} \sum_1^n (X_i - \mu)^2}_A = \underbrace{\frac{1}{\sigma^2} \sum_1^n (X_i - \bar{X})^2}_B + \underbrace{\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \right)^2}_C$$

We have:

$$A = \sum_1^n \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi_n^2$$

$$B \sim ?$$

$$C \sim \chi_1^2$$

So we argue as follows:

$$M_A(t) = \frac{1}{(1-2t)^{\frac{n}{2}}}$$

$$M_C(t) = \frac{1}{(1-2t)^{\frac{1}{2}}}$$

Hopefully, we will be able to show independence (done in Theorem 1.12), we will have:

$M_A(t) = M_{B+C}(t) = M_B(t) \cdot M_C(t)$ since then:

$$M_B(t) = \frac{M_A(t)}{M_C(t)} = \frac{1}{(1-2t)^{\frac{n-1}{2}}}$$

and by uniqueness, finally,

$$\frac{(n-1)s^2}{\sigma^2} = B \sim \chi_{n-1}^2$$

□

Theorem 1.12. *Let $X_1, X_2, \dots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. Then s^2 and $\bar{X}$ are independent.*

Proof. First, without loss of generality, we assume that the X_i 's are i.i.d. $N(0, 1)$.

Now,

$$\begin{aligned} (n-1)s^2 &= \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \sum_{i=2}^n (X_i - \bar{X})^2 + (X_1 - \bar{X})^2 \\ &= \sum_{i=2}^n (X_i - \bar{X})^2 + \left[\sum_{i=2}^n (X_i - \bar{X}) \right]^2 \end{aligned}$$

Reason:

$$\begin{aligned} \sum_{i=1}^n (X_i - \bar{X}) &= 0 \text{ by definition of the sample mean} \\ \Rightarrow (X_1 - \bar{X}) &= - \sum_{i=2}^n (X_i - \bar{X}) \\ \Rightarrow \left[(X_1 - \bar{X}) \right]^2 &= \left[\sum_{i=2}^n (X_i - \bar{X}) \right]^2 \end{aligned}$$

Therefore, miraculously, we can write $(n-1)s^2 = \sum_{i=1}^n (X_i - \bar{X})^2$ as a function, g , say of $(X_2 - \bar{X}), (X_3 - \bar{X}), \dots, (X_n - \bar{X})$

Hence, it will be sufficient if we can show that $\bar{X}$ is independent of $(X_2 - \bar{X}), (X_3 - \bar{X}), \dots, (X_n - \bar{X})$

To this end, make the following multivariate transformation from $(x_1, \dots, x_n)$ to $(y_1, \dots, y_n)$ as follows:

$$\begin{array}{ll}
Y_1 = \bar{X} & \\
Y_2 = X_2 - \bar{X} & X_2 = Y_2 + Y_1 \\
\vdots & \vdots \\
Y_n = X_n - \bar{X} & X_n = Y_n + Y_1
\end{array}$$

With the defined transformation, it is easy to show that $|J| = |\det\{\frac{\partial X_i}{\partial y_i}\}| = n$ and that all the conditions of the transformation theorem are satisfied. The transformation, in this case, happens to be linear.

Therefore,

$$f_{Y_1, Y_2, \dots, Y_n}(y_1, \dots, y_n) = f_{X_1, X_2, \dots, X_n}(y_1, \dots, y_n) \cdot |J|$$

Now:

$$f_{X_1, X_2, \dots, X_n}(x_1, \dots, x_n) = \frac{1}{(2\pi)^{\frac{n}{2}}} \cdot e^{-\frac{1}{2} \sum_1^n x_i^2}$$

Since $X_i \stackrel{i.i.d.}{\sim} N(0, 1)$ and we note that $\sum_1^n x_i^2 = \sum_1^n (x_i - \bar{X})^2 + n\bar{X}^2$

Therefore,

$$f_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = \frac{1}{(2\pi)^{\frac{n}{2}}} \cdot e^{-\frac{1}{2} \left[\sum_2^n y_i^2 + \left[\sum_2^n y_i \right]^2 + n y_1^2 \right]} \cdot n$$

And since we can factor out a function of y_1 , we are done.

Now, we invoke the following theorem from assignment 2:

$X_1, \dots, X_n$ are independent if and only if $f_{X_1, \dots, X_n}(x_1, \dots, x_n)$ can be written as $g_1(x_1), \dots, g_n(x_n)$

So this implies:

$$\begin{aligned}
& (Y_2, Y_3, \dots, Y_n) \perp Y_1 \\
& \text{i.e. } \bar{X} \perp (X_2 - \bar{X}, \dots, X_n - \bar{X})
\end{aligned}$$

Therefore, we get that $\bar{X} \perp s^2$

□

Now, we are finally ready to prove the original theorem.

Proof. (Of Theorem 1.9)

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \stackrel{(\star)}{=} \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \bigg/ \sqrt{\frac{(n-1)s^2}{\sigma^2}} \frac{1}{n-1}$$

($\star$) is by algebraic manipulation

But by Theorem 1.11, the numerator of the denominator, $\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$. And we know that $\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$. So using Theorem 1.12, which gives us independence, we get, by Theorem 1.10, the intended result.

So therefore, after a long and horrible ordeal,

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

□

Theorem 1.13. Let $X_1, X_2, \dots, X_{n_1}$ be i.i.d. $N(\mu_1, \sigma^2)$ and $Y_1, Y_2, \dots, Y_{n_2}$ be i.i.d. $N(\mu_2, \sigma^2)$

Further, let $\tilde{X} \perp \tilde{Y}$

$$\begin{aligned} \text{Let } s_X^2 &= \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2 \\ s_Y^2 &= \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2 \end{aligned}$$

Then,

$$\frac{s_x^2}{s_y^2} \sim F_{n_1-1, n_2-1}$$

[Aside: the F-distribution has support on $(0, \infty)$ and is skewed to the right. It depends on two parameters, n_1 and n_2 called the degrees of freedom]

Proof.

$$\frac{s_x^2}{s_y^2} = \frac{\frac{(n_1-1)s_x^2}{\sigma^2}}{n_1 - 1} \bigg/ \frac{\frac{(n_2-1)s_y^2}{\sigma^2}}{n_2 - 1}$$

And we note that:

$$\frac{(n_1 - 1)s_x^2}{\sigma^2} \sim \chi_{n_1-1}^2$$

$$\frac{(n_2 - 1)s_y^2}{\sigma^2} \sim \chi_{n_2-1}^2$$

And since $\frac{(n_1-1)s_x^2}{\sigma^2} \perp \frac{(n_2-1)s_y^2}{\sigma^2}$ we can apply Question 2 from Assignment 1, and we're done.

□

2 Statistical Inference

2.1 Introduction

Suppose that we have a population whose values are determined by some probability distribution $F_X(x; \theta)$, where θ is a parameter, possibly vector-valued, that indexes the members of the parametric family of distributions assumed to define the population.

If at least one value of θ , i.e. one component, is unknown, then we have a parametric statistical inference problem.

Example: If X_i are i.i.d. $N(\mu, 3.6)$ with μ unknown, then the inference problem concerns the mean, μ . If X_i are i.i.d. $N(\mu, \sigma^2)$ with both μ and σ^2 unknown, then the inference problem is about $(\mu, \sigma^2) = \theta$

To make our inference, we will draw a random sample $X_1, \dots, X_n$ of i.i.d. observations $X_i \sim F_{X_1}(x; \theta)$.

We have two main types of problems:

1. Guess at θ . [Estimating θ]
2. Decide which of two sets, Θ_1 or Θ_2 , θ belongs to. [Hypothesis testing]

Definition: Any (measurable) function of the observation, $T = T(X)$, which is independent of unknown parameters, is called a *statistic*.

Examples:

If $(X_1, X_2, \dots, X_n)$ have unknown mean μ , then the following are statistics:

1. $T_1 = 73X_{25}$
2. $T_2 = \frac{X_1 + X_5}{2}$
3. $T_3 = \sqrt{X_{10}}$
4. $T_4 = \bar{X}$
5. $T_5 = 25$

but $T_6 = \mu + X_4$ is not a statistic.

Defintion: Any statistic, T , whose purpose is to estimate at θ is called an *estimator*. We denote estimators by $\hat{\theta}$, which is a random variable itself, and therefore has its own distribution $f_{\hat{\theta}}$, which is called the sampling distribution of $\hat{\theta}$.

Suppose now that the experiment is completed so that we have observed or realized values of $X_1, \dots, X_n$. Then $\hat{\theta}(x_1, \dots, x_n)$ is called an *estimate*, and is a non-random, post-experiment quantity.

Examples:

1. Let $X_1, \dots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ with both μ and σ^2 unknown. Then the following are estimators:
 - (a) $\hat{\mu} = \frac{X_1 + X_2}{2}$
 - (b) $\hat{\mu} = X_1$
 - (c) $\hat{\mu} = \bar{X}$
 - (d) $\hat{\sigma}^2 = X_3^2 + X_7^2$
 - (e) $\hat{\sigma}^2 = s^2$
2. Suppose $X_1, \dots, X_n$ are i.i.d. $\text{Bern}(p)$ random variables, then $\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$ is an estimator for p , as is $\hat{p} = X_3$
3. Suppose $X_1, \dots, X_n$ are i.i.d. $U(0, \theta)$, with θ unknown. Then $\hat{\theta} = Y_n$ is an estimator, or $\hat{\theta} = 2\bar{X}$

The aim is to find “good” estimators, and in some cases, the “best” estimators. To do this, we need to define criteria of “good-ness.”

2.2 Properties of Estimators

In our search for good estimators, we might seek estimators whose distribution $f_{\hat{\theta}}(y)$ is closely concentrated about the unknown parameter θ .

Note that since θ is unknown, we can say nothing about how close a $\hat{\theta}(x_1, \dots, x_n)$ observed, is to θ . We can only talk about the probability that $\hat{\theta}$ will be close before the experiment. Once the experiment has been completed, we can only be secure in the knowledge that over a large number of trials our estimator is good.

This discussion leads to the first desirable property for an estimator.

2.2.1 Unbiasedness and Minimum Variance

Definition: An estimator $\hat{\theta}$ is said to be *unbiased* for θ if

$$\mathbb{E}(\hat{\theta}) \equiv \theta, \forall \theta \in \Theta \text{ (where } \Theta \text{ is the set of possible values for } \theta.)$$

Note:

1. The equality must hold for all θ . Thus it is no good if $\mathbb{E}_{\theta}(\hat{\theta}) = \theta_0$ for $\theta = \theta_0$.
2. Heuristically, an estimator is unbiased if on average, over a large number of repeated uses, it is equal to θ
3. Reason for terminology: If $\mathbb{E}(\hat{\theta}) \neq \theta$ then we call $\hat{\theta}$ biased for θ and we call $\mathbb{E}(\hat{\theta}) - \theta$ the bias. For an unbiased estimator, the $\hat{\theta}$ fall evenly on either side of θ , but for a biased estimator, the values of $\hat{\theta}$ tend to more often fall to one side of θ , but will still fall evenly about $\mathbb{E}(\hat{\theta})$.
4. When the sampling distribution is skewed, unbiasedness is not useful.
5. Of course we could talk of some function, $g(\theta)$ and $g(\hat{\theta})$ being unbiased for $g(\theta)$

Examples:

1. $\bar{X}$ is always unbiased for μ , since $\mathbb{E}(\bar{X}) \stackrel{\mu}{=} \mu$.
2. Claim: $\mathbb{E}(s^2) \stackrel{\sigma^2}{=} \sigma^2$, for all distributions of X_i .

Proof. Note that:

$$\sum_1^n (X_i - \bar{X})^2 = \sum_1^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

Hence:

$$\begin{aligned} \mathbb{E}(s^2) &= \frac{1}{n-1} \left[\sum_1^n \mathbb{E}[(X_i - \mu)^2] - n\mathbb{E}[(\bar{X} - \mu)^2] \right] \\ &= \frac{1}{n-1} \left[n\sigma^2 - n \cdot \frac{\sigma^2}{n} \right] \\ &= \sigma^2 \end{aligned}$$

□

3. Let $X_1, \dots, X_n$ be i.i.d. Bern(p) then $\hat{p} = \frac{\bar{X}}{n} = \frac{1}{n} \sum X_i = \bar{X}$, the sample proportion, is unbiased for p .
4. Let $X_1, \dots, X_n$ be i.i.d. $U(0, \theta)$ random variables. Let $\hat{\theta}_1 = Y_n =$ the largest order statistic. Let $\hat{\theta}_2 = 2\bar{X}$. First, $\hat{\theta}_2$ is unbiased for θ since

$$\mathbb{E}(\hat{\theta}_2) = 2 \cdot \mathbb{E}(\bar{X}) = 2 \cdot \frac{\theta}{2} = \theta$$

To find $\mathbb{E}(\hat{\theta}_1)$, we need $f_{Y_n}(y)$.

We have:

$$\begin{aligned} F_{Y_n}(y) &= P(X_1 \leq y)P(X_2 \leq y) \cdots P(X_n \leq y) \\ &= [F_{X_1}(y)]^n \\ &= \begin{cases} \left(\frac{y}{\theta}\right)^n & 0 < y < \theta \\ 1 & y \geq \theta \\ 0 & y \leq 0 \end{cases} \end{aligned}$$

$$\text{Thus, } f_{Y_n}(y) = \begin{cases} \frac{ny^{n-1}}{\theta^n} & 0 < y < \theta \\ 0 & \text{else} \end{cases}$$

So therefore,

$$\begin{aligned} \mathbb{E}(Y_n) &= \int_0^\theta y \cdot \frac{ny^{n-1}}{\theta^n} dy \\ &= n \cdot \frac{\theta^{n+1}}{(n+1)\theta^n} \\ &= \frac{n}{n+1} \cdot \theta \\ &< \theta \text{ for all } n \end{aligned}$$

which shows that Y_n is biased for θ . However, $\frac{n+1}{n} \cdot Y_n$ is unbiased for θ . We shall need to choose between $\frac{n+1}{n} \cdot Y_n$ and $2\bar{X}$ as estimators for θ .

5. X_1 is unbiased for μ since $\mathbb{E}(X_1) = \mu$ and thus we shall have to choose between $\hat{\theta}_1 = \bar{X}$ and $\hat{\theta}_2 = X_1$

We realize that unbiasedness alone is not sufficient to make an estimator good. Clearly, X_1 and $\bar{X}$ are both unbiased estimators for μ , but X_1 is, on intuitive grounds, bad. Roughly, X_1 is much more variable than $\bar{X}$ is. In fact, $Var(X_1) = \sigma^2$ while $Var(\bar{X}) = \frac{\sigma^2}{n}$.

We are thus lead to the consideration of variance as a criteria of goodness for an estimator. If possible, we seek estimators that have smaller variance than any other unbiased estimators.

Definition: $\hat{\theta}$ is called a *minimum variance unbiased estimator* for θ if:

1. $\mathbb{E}(\hat{\theta}) \stackrel{\theta}{=} \theta$
2. If $\hat{\theta}^*$ is any other estimator for θ , such that $\mathbb{E}(\hat{\theta}^*) \stackrel{\theta}{=} \theta$, then

$$Var_\theta(\hat{\theta}) \leq Var_\theta(\hat{\theta}^*) \quad \forall \theta \in \Theta$$

We say that $\hat{\theta}$ is an MVUE or UMVUE (uniform minimum variance unbiased estimator).

The Cramér-Rao Inequality

The following inequality on the variance of an unbiased estimator allows us, in some situations, to determine whether a given estimator is MVU.

Theorem 2.1. *Let $\underline{X}$ have p.d.f. $f_{\underline{X}}(\underline{x}; \theta)$ where θ is scalar valued. Let $T(\underline{X})$ be an unbiased estimator for $g(\theta)$. Let $\text{Var}\left[\frac{\partial}{\partial\theta}f_{\underline{X}}(\underline{x}; \theta)\right]$ be finite, and suppose that*

$$\int \frac{\partial}{\partial\theta}T(\underline{X})f_{\underline{X}}(\underline{x}; \theta)d\underline{x} = \frac{\partial}{\partial\theta} \int T(\underline{X})f_{\underline{X}}(\underline{x}; \theta)d\underline{x}$$

Then $\text{Var}(T(\underline{X})) \geq \frac{[g'(\theta)]^2}{\mathbb{E}[\frac{\partial}{\partial\theta}\ln f_{\underline{X}}(\underline{x}; \theta)]^2}$ where the right hand side is called the Cramér-Rao Lower Bound.

Proof. First, we need the following lemma

Lemma 2.2. $\mathbb{E}[\frac{\partial}{\partial\theta}\ln f_{\underline{X}}(\underline{x}; \theta)] = 0$

Proof. (of lemma)

$$\begin{aligned}\mathbb{E}[\frac{\partial}{\partial\theta}\ln f_{\underline{X}}(\underline{x}; \theta)] &= \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta}\ln f_{\underline{X}}(\underline{x}; \theta) \cdot f_{\underline{X}}(\underline{x}; \theta)d\underline{x} \\ &= \frac{\partial}{\partial\theta} \int_{-\infty}^{\infty} f_{\underline{X}}(\underline{x}; \theta)d\underline{x} \\ &= \frac{\partial}{\partial\theta}(1) \\ &= 0\end{aligned}$$

□

Now, back to the main proof.

$$\underbrace{\left[Cov\left(T(\underline{X}), \frac{\partial}{\partial\theta}\ln f_{\underline{X}}(\underline{x}; \theta)\right)\right]^2}_A \leq \text{Var}(T(\underline{X})) \cdot \text{Var}\left[\frac{\partial}{\partial\theta}\ln f_{\underline{X}}(\underline{x}; \theta)\right]$$

(By Cauchy-Schwartz, or by noting that $-1 \leq \rho \leq 1$ and that $\rho = \frac{Cov}{Var}$)

Note that, by the lemma:

$$Var[\frac{\partial}{\partial\theta} \ln f_{\underline{X}}(\underline{x}; \theta)] = \mathbb{E}[(\frac{\partial}{\partial\theta} \ln f_{\underline{X}}(\underline{x}; \theta))^2] - \underbrace{\left(\mathbb{E}(\frac{\partial}{\partial\theta} \ln f_{\underline{X}}(\underline{x}; \theta))\right)^2}_{=0} = \mathbb{E}[(\frac{\partial}{\partial\theta} \ln f_{\underline{X}}(\underline{x}; \theta))^2]$$

Thus, consider A:

$$\begin{aligned} A &= [\mathbb{E}[T(\underline{X}) \frac{\partial}{\partial\theta} \ln f_{\underline{X}}(\underline{x}; \theta)] - \underbrace{\mathbb{E}(T(\underline{X})) \mathbb{E}[\frac{\partial}{\partial\theta} \ln f_{\underline{X}}(\underline{x}; \theta)]}_{=0}]^2 \\ &= [\int_{-\infty}^{\infty} T(\underline{X}) [\frac{\partial}{\partial\theta} \ln f_{\underline{X}}(\underline{x}; \theta)] f_{\underline{X}}(\underline{x}; \theta) dx]^2 \\ &= [\int_{-\infty}^{\infty} T(\underline{X}) \frac{\partial}{\partial\theta} f_{\underline{X}}(\underline{x}; \theta) dx]^2 \\ &= [\frac{\partial}{\partial\theta} \int_{-\infty}^{\infty} T(\underline{X}) f_{\underline{X}}(\underline{x}; \theta) dx]^2 \\ &= [\frac{\partial}{\partial\theta} \mathbb{E}[T(\underline{X})]]^2 \\ &= [g'(\theta)]^2 \end{aligned}$$

And hence we can conclude our result. □

Note: If $g(\theta) = \theta$ then $[g'(\theta)]^2 = 1$. Also, we have equality if and only if $T(\underline{X}) = a(\theta) \frac{\partial}{\partial\theta} \ln f_{\underline{X}}(\underline{x}; \theta) + b(\theta)$ for constants $a(\theta)$ and $b(\theta)$.

Corollary 2.3. Suppose in the C-R inequality, $X_1, X_2, \dots, X_n$ are i.i.d. with p.d.f. $f_X(x; \theta)$.

Then

$$Var(T(X)) \geq \frac{[g'(\theta)]^2}{n \mathbb{E}[(\frac{\partial}{\partial\theta} \ln f_X(x; \theta))^2]}$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{d}{d\theta} \ln f_{\underline{X}}(\underline{x}; \theta)\right)^2\right] &= \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \ln \prod f_{X_i}(x_i; \theta)\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{d}{d\theta} \sum_{i=1}^n \ln f_{X_i}(x_i; \theta)\right)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^n \left(\frac{d}{d\theta} \ln f_{X_i}(x_i; \theta)\right)^2\right] + \mathbb{E}\left[\frac{d}{d\theta} \sum_{i \neq j} \ln f_{X_i}(x_i; \theta) \ln f_{X_j}(x_j; \theta)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^n \left(\frac{d}{d\theta} \ln f_{X_i}(x_i; \theta)\right)^2\right] + \underbrace{\mathbb{E}\left[\sum_{i \neq j} \left(\frac{\partial}{\partial\theta} \ln f_{X_i}(x_i; \theta)\right) \left(\frac{\partial}{\partial\theta} \ln f_{X_j}(x_j; \theta)\right)\right]}_{=0 \text{ by lemma}} \\
&= \sum_{i=1}^n \mathbb{E}\left[\left(\frac{d}{d\theta} \ln f_{X_i}(x_i; \theta)\right)^2\right] \\
&= n \mathbb{E}\left[\left(\frac{d}{d\theta} \ln f_{X_1}(x_1; \theta)\right)^2\right]
\end{aligned}$$

□

Example: Let $X_1, \dots, X_n$ be i.i.d. Bernoulli random variables with parameter, p , unknown. Show that $\hat{p} = \frac{\underline{X}}{n} \stackrel{\text{def}}{=} \frac{1}{n} \sum X_i$, is a MVUE for p .

Solution: We will try to show that $\text{Var}(\hat{p}) = \text{C-R Lower Bound}$, for then, $\hat{p}$ must be MVU.

First, $\text{Var}(\hat{p}) = \text{Var}\left(\frac{\underline{X}}{n}\right) = \frac{1}{n^2} \cdot np(1-p) = \frac{p(1-p)}{n}$, so we must show that the C-R lower bound is $\frac{p(1-p)}{n}$. Since the X_i 's are i.i.d., the C-R lower bound is:

$$\frac{1}{n \mathbb{E}\left[\left(\frac{\partial}{\partial p} \ln f_{\underline{X}}(\underline{x}; p)\right)^2\right]} = \frac{1}{n \mathbb{E}\left[\left(\frac{\partial}{\partial p} x \ln(p) + (1-x) \ln(1-p)\right)^2\right]} = \frac{(p(1-p))^2}{n \mathbb{E}[(X-p)^2]} = \frac{p(1-p)}{n}$$

And so, the C-R lower bound is equal to the variance, and therefore $\hat{p}$ is MVU.

Theorem 2.4. *If $T(\underline{X})$ is MVU for $g(\theta)$, then it is unique.*

Proof. Suppose that $T_1(\underline{X})$ is MVU for $g(\theta)$ and let $T_2(\underline{X})$ be another MVUE for $g(\theta)$. Define $T(\underline{X}) = \frac{T_1(\underline{X}) + T_2(\underline{X})}{2}$. We note that $T(\underline{X})$ is also unbiased for $g(\theta)$.

Consider:

$$\begin{aligned}
Var(T(\underline{X})) &= \frac{1}{4}(Var(T_1 + T_2)) \\
&= \frac{1}{4}(Var(T_1) + Var(T_2) + 2Cov(T_1, T_2)) \\
&\leq \frac{1}{4}(2Var(T_1) + 2Var(T_1)) \\
&= Var(T_1)
\end{aligned}$$

If we have a strict inequality, then we reach a contradiction to the fact that T_1 is MVUE.

If $Var(T) = Var(T_1)$, this happens if and only if $T_1(X) = a(\theta)T_2(X) + b(\theta)$. Because of unbiasedness, we get that $b(\theta) = 0$ and $Var(T_1) = a^2(\theta)Var(T_2)$ gives that $a(\theta) = 1$

Because $\rho^2 \leq 1$ and $Var(T_1) = Var(T_2)$. Furthermore, this is a contradiction because T_1 is supposed to be MVUE. $\square$

Remark: The C-R inequality is useful in showing that an estimator is MVU. You have to be lucky, though, since if $Var(T) >$ C-R lower bound, then nothing can be said about the “MVU-ness” of T. In addition, the C-R lower bound does not give a way of finding an MVUE.

2.2.2 Consistency

Another desirable property for an estimator, $\hat{\theta}$ is that it should “approach” the parameter θ as the sample size, n , tends to infinity. This leads to the following definition.

Definition: $\hat{\theta}$ is said to be *weakly consistent* for θ if $\hat{\theta} \xrightarrow{P} \theta$ as $n \rightarrow \infty$. ($\hat{\theta}$ is called *strongly consistent* for θ if $\hat{\theta} \xrightarrow{WP1} \theta$ as $n \rightarrow \infty$)

Ways of Determining Weak Consistency:

1. Use the definition of consistency directly
2. If $\hat{\theta}$ can be written as an average of i.i.d. random variables, use the Weak Law of Large Numbers.

3. If $Var(\hat{\theta}) \rightarrow 0$ as $n \rightarrow \infty$ and $\mathbb{E}(\hat{\theta}) \rightarrow \theta$ as $n \rightarrow \infty$, then $\hat{\theta} \xrightarrow{P} \theta$

Definition: If $\mathbb{E}_{\theta}(\hat{\theta}) \rightarrow \theta$ as $n \rightarrow \infty$ for all θ , then $\hat{\theta}$ is said to be *asymptotically unbiased* for θ .

Proof. (of 3) If $\mathbb{E}(\hat{\theta}) \equiv \theta$, then Markov's Inequality would ensure that $\hat{\theta} \xrightarrow{P} \theta$.

$$\begin{aligned} P(|\hat{\theta} - \theta| > \delta) &= P(|\hat{\theta} - \mathbb{E}(\hat{\theta})| > \delta) \\ &\leq \frac{\mathbb{E}(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2}{\delta^2} \\ &= \frac{Var(\hat{\theta})}{\delta^2} \xrightarrow{n \rightarrow \infty} 0 \end{aligned}$$

On the other hand, with asymptotic unbiasedness, just note that:

$$\hat{\theta} - \theta = \hat{\theta} - \mathbb{E}(\hat{\theta}) + \mathbb{E}(\hat{\theta}) - \theta = \underbrace{\hat{\theta} - \mathbb{E}(\hat{\theta})}_{\text{by Markov}} + \underbrace{\mathbb{E}(\hat{\theta} - \theta)}_{\text{asymptotically unbiased}}$$

where both terms of the right hand side approach 0 as $n \rightarrow \infty$ □

Examples:

1. Let $X_1, X_2, \dots, X_n$ be i.i.d. Bern(p) random variables, then, by the Weak Law of Large Numbers, as $n \rightarrow \infty$:

$$\hat{p} \stackrel{def}{=} \frac{X}{n} \equiv \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P} p$$

and hence $\hat{p}$ is consistent for p .

2. $\bar{X}$ is always consistent for $\mu = \mathbb{E}(X_1)$ for i.i.d. random variables $X_1, \dots, X_n$ with finite mean.
3. s^2 is always strongly consistent for σ^2 provided the X_i 's are i.i.d. with finite variance. (For proof, see Theorem 1.7)
4. $\hat{\mu}_1 \stackrel{def}{=} \bar{X} + \frac{3}{\log(n)}$ is consistent for μ , since $\hat{\mu}_1$ is asymptotically unbiased for μ and hence, $Var \rightarrow 0$ gives the result, or more simply, we notice that $\bar{X} \xrightarrow{P} \mu$ and $\frac{3}{\log(n)} \rightarrow 0$ as $n \rightarrow \infty$.

5. Let $X_1, \dots, X_n$ be i.i.d. $U(0, \theta)$. Define $\hat{\theta} = \frac{n+1}{n}Y_n$. Then, $\hat{\theta}$ is consistent for θ as, indeed, is Y_n .

Proof. We could use Markov's Inequality, but instead, we directly consider $\hat{\theta}_1 = Y_n$

$$\begin{aligned} P(|\hat{\theta}_1 - \theta| > \delta) &= P(|Y_n - \theta| > \delta) \\ &= P(\theta - Y_n > \delta) \text{ for } \theta > Y_n \\ &= P(Y_n < \theta - \delta) \quad (\star) \end{aligned}$$

Now since the support of the X_i 's is $(0, \theta)$, $P(Y_n < y)$ for $y < 0$ is 0. Therefore, if $\delta > \theta$, then $P(Y_n < \theta - \delta) = 0$, so consider $\delta < \theta$. Then,

$$(\star) = \left(\frac{\theta - \delta}{\theta} \right)^n \xrightarrow{n \rightarrow \infty} 0$$

And so, we're done, since $\hat{\theta}_1 \xrightarrow{P} \theta$ as $n \rightarrow \infty$ and so does $\hat{\theta}$ □

2.2.3 Sufficiency

Much of statistics is concerned with summarizing or reducing data. We seek a statistic, T , say, that summarizes the data in the sense that once T is known, we can dispense with the original data in making inference about θ

We make this idea concrete through the following definition. Henceforth, unless said otherwise, f_X will denote either a p.d.f. or a p.m.f.

Definition: Let $X_1, X_2, \dots, X_n = \underline{X} \sim f_{\underline{X}}(x; \theta)$ where θ could be vector-valued. We say that $T = T(\underline{X})$ is *sufficient* for θ , if, $f_{\underline{X}; \theta|T=t}(\underline{X}|t)$ is mathematically independent of θ for all $\underline{X}$, for all t such that $f_T(t) \neq 0$.

Generating Sufficient Statistics:

Suppose that $X_1, \dots, X_n = \underline{X} \sim f_{\underline{X}}(x; \theta)$. Suppose that $T = T(\underline{X})$ is observed and that θ is unknown. Now suppose that your dog eats your data, $\underline{X}$, but you still have T , where T is sufficient for θ .

Is it possible to generate a sample, $\underline{X}^* = X_1^*, \dots, X_n^*$, which has the same distribution as $\underline{X}$?

Note that since θ is unknown, even though you have $f_{\underline{X}}(\underline{x}; \theta)$, you are unable to simulate an $\underline{X}^*$ from $f_{\underline{X}}$. We use the following means of generating a pair of random variables with joint density (pmf) $f_{X,Y}(x, y)$:

1. Generate y^* from $f_Y(y)$
2. Generate x^* from $f_{X|Y=y}(x|y)$.

We now have the pair, (x^*, y^*) and we claim that $(X^*, Y^*) \sim f_{X,Y}(x, y)$. Let $Y^* = T$ and we're done.

Notes:

1. Sufficiency plays an important role in statistical inference since it often enables one to work with the distribution of a simpler random variable, T , than with the random variables $X_1, \dots, X_n$ themselves. Moreover, as we will see, surprisingly, sufficiency plays a role in finding MVUE's.
2. A sufficient statistic may be vector valued - even for a scalar parameter. The trivial sufficient statistic is $T(\underline{X}) = (X_1, X_2, \dots, X_n)$ since $f_{\underline{X}}(\underline{x}|\underline{X} = \underline{x}) = 1$ independent of θ for all $\underline{x}$ such that $f_{\underline{X}}(\underline{x}) = 0$
3. Any one-to-one function of a sufficient statistic is sufficient.
4. A sufficient statistic may not be a good estimator without modification.

For example, we'll see that $T = \sum_{i=1}^n X_i$ is sufficient for p , in a Bernoulli setting (hence, so is $T^* = \frac{1}{n} \sum_{i=1}^n X_i$) but T is clearly unsuitable for estimation of p .

5. We can look at sufficiency slightly differently. If T is given, it induces a partition of the sample space, S , as follows:

Let $A_t = \{\underline{x} : T(\underline{x}) = t\}$.

Clearly, $\bigcup_t A_t = S$ and $A_s \cap A_t = \emptyset$ for all $s \neq t$

We call $\mathcal{A}_T = \{A_t, t \in I\}$ a sufficient partition of S if T is sufficient. The partition is sufficient if given $\underline{X} \in A_t$, $f(\underline{x})$ is independent of θ for any $\underline{x} \in A_t$

6. In general, the sufficiency of T will depend on the family of distributions from which data arise.
7. $f_T(t)$ will, of course, depend on θ
8. The original data should not be discarded in case you need to check the model assumptions, which are needed for sufficiency.

Example: Let $X_1, \dots, X_n$ be i.i.d. Bernoulli(p) random variables with p unknown. Prove that $T(\underline{X}) = \sum_{i=1}^n X_i =: X$ is sufficient for p . That is, given the number of successes in n trials, there is no more useful information about p in the data.

Proof. We must show that $P(\underline{X} = \underline{x} | X = x)$ is independent of p .

First, note that $P(\underline{X} = \underline{x} | X = x) = 0$ if $x \neq \sum_{i=1}^n x_i$, which is independent of θ .

So we have:

$$\begin{aligned}
 P(\underline{X} = \underline{x} | X = \sum_{i=1}^n x_i) &\stackrel{A \subseteq B}{=} \frac{P(\underline{X} = \underline{x})}{P(X = \sum_{i=1}^n x_i)} \\
 &= \frac{p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}}{\left(\sum_{i=1}^n x_i \right) p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}} \\
 &= \frac{1}{\binom{n}{\sum_{i=1}^n x_i}} \text{ which is independent of } p
 \end{aligned}$$

□

It turns out that one may avoid showing sufficiency by finding conditional distributions. Further, one can avoid needing to have a T in mind.

Theorem 2.5. (*Neyman Factorization Theorem*)

Let $\underline{X} \sim f_{\underline{X}}(\underline{x}; \theta)$, where θ is unknown. Then T is sufficient for θ if and only if $f_{\underline{X}}(\underline{x}; \theta) = g(T(\underline{x}); \theta)h(\underline{x})$ where g is a function of θ and of the data $\underline{x}$ through T alone, and where $h(\underline{x})$ is a function of the data only.

Proof. We give the proof in the discrete case only (the continuous case is too difficult).

First, suppose that T is sufficient according to the definition. We must show that the density factors:

$$P(\underline{X} = \underline{x}; \theta) = \overbrace{P(\underline{X} = \underline{x} | T(\underline{X}) = t)}^{h(\underline{x})} \cdot \overbrace{P(T(\underline{X}) = t)}^{g(T(\underline{x}); \theta)} + \overbrace{P(\underline{X} = \underline{x} | T(\underline{X}) \neq t)}^A \cdot P(T(\underline{X}) \neq t)$$

Clearly, $A=0$ if $t \neq T(\underline{x})$ and so we're done.

Conversely, suppose that f factors. Now we must show that the conditional probability is independent of θ .

$$\begin{aligned} P(\underline{X} = \underline{x} | T(\underline{X}) = T(\underline{x})) &= \frac{P(\underline{X} = \underline{x}; \theta)}{P(T(\underline{X}) = T(\underline{x}))} \\ &= \frac{g(T(\underline{x}); \theta)h(\underline{x})}{\sum_{\underline{y}: T(\underline{y})=T(\underline{x})} P(\underline{X} = \underline{y})} \\ &= \frac{g(T(\underline{x}); \theta)h(\underline{x})}{\sum_{\underline{y}: T(\underline{y})=T(\underline{x})} h(\underline{y})g(T(\underline{y}); \theta)} \\ &= \frac{h(\underline{x})}{\sum_{\underline{y}: T(\underline{y})=T(\underline{x})} h(\underline{y})} \end{aligned}$$

which is independent of θ . □

Example: Let $X_1, \dots, X_n$ be i.i.d. $U(0, \theta)$ random variables with θ unknown. Find a sufficient statistic for θ .

Solution: We have:

$$\begin{aligned} f_{\underline{X}}(\underline{x}) &= \frac{1}{\theta^n} \prod_{i=1}^n \mathbb{I}_{(0,\theta)}(x_i) \\ &= \underbrace{\frac{1}{\theta^n} \mathbb{I}_{(0,\theta)}(\max(x_i) < \theta)}_g \cdot \underbrace{\mathbb{I}_{(0,\infty)}(\min(x_i) > 0)}_h \end{aligned}$$

So therefore, $T(\underline{X}) = \max(X_1, \dots, X_n)$ is sufficient for θ .

Example: Let $X_1, \dots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ random variables where μ and σ^2 are both unknown. Find a sufficient statistic for $\theta := (\mu, \sigma^2)$

Solution:

$$\begin{aligned} f_{\underline{X}}(\underline{x}; \mu, \sigma^2) &= \frac{1}{(2\pi)^{\frac{n}{2}}} \cdot \frac{1}{\sigma^n} e^{-\frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^2}} \cdot \prod_{i=1}^n \mathbb{I}_{(-\infty, \infty)}(x_i) \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{\sigma^n} e^{-\frac{1}{2} \sum_{i=1}^n X_i^2 - 2\mu \sum_{i=1}^n X_i + n\mu^2} \cdot \prod_{i=1}^n \mathbb{I}_{(-\infty, \infty)}(x_i) \end{aligned}$$

so $T(\underline{X}) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is sufficient for (μ, σ^2)

Comment: Since $s^2 = \frac{1}{n-1} [\sum_{i=1}^n (X_i - \bar{X})^2] = \frac{1}{n-1} [\sum_{i=1}^n X_i^2 - n\bar{X}^2]$, it immediately follows

that $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is a one-to-one function of $(\bar{X}, s^2)$. So $(\bar{X}, s^2)$ is also sufficient for (μ, σ^2) and its components are sensible estimators for μ and σ^2 respectively.

The following question immediately arises. Once we have identified a sufficient statistic, is it possible to “reduce” the information about θ in this statistic and still retain sufficiency. For example, $\underline{X}$ is always sufficient for any parameter, but in some cases it does not give the greatest possible reduction. This leads to:

Definition: A statistic, T , is called *minimal sufficient* for θ if:

1. It is sufficient for θ
2. If T^* is any other sufficient statistic for θ , then T is a function of T^* . That is, there exists a g such that $T = g(T^*)$.

In terms of partitions, it is easy to show that a minimal sufficient statistic induces the coarsest partition of S amongst all sufficient partitions. That is, any other sufficient partition is a sub-partition of that induced by T .

Beware, it does not follow that if you have a 1-dimensional sufficient statistic that it must be minimal.

Exercise: Find an example of a 1-dimensional sufficient statistic (for a scalar-valued) parameter that is not minimal sufficient.

Solution: (Ioan's Solution) Let $X_0, X_1, \dots, X_n$ be i.i.d. $Bern(p)$ random variables. Let $T = \sum_{i=0}^n 10^i X_i$. There is a one-to-one correspondence between the X_i 's and T , and therefore

it's sufficient. This statistic, though, is not minimal. We know that $X = \sum_{i=0}^n X_i$ is sufficient for p . If T was minimal, we would be able to write $T = f(X)$. We consider $x_1 = (1, 0, 0, \dots, 0)$ and $x_2 = (0, 1, 0, \dots, 0)$. Since $X(x_1) = X(x_2) = 1$, we would need that $T(x_1) = T(x_2)$, but this is not true.

Comment: It is easy to find an example of a scalar valued parameter, whose minimal sufficient statistic is of higher dimensions.

The following allows us to find a minimal sufficient statistic which, initially, would seem to be hard.

Lemma 2.6. Let $\underline{X} \sim f_{\underline{X}}(\underline{x}; \theta)$. Let T be sufficient for θ . Suppose that for $\underline{x}, \underline{y}$, we have $T(\underline{x}) = T(\underline{y})$. Then, $\frac{f(\underline{x}; \theta)}{f(\underline{y}; \theta)}$ is independent of θ .

Proof. By the Neyman Factorization Theorem,

$$\frac{f(\underline{x}; \theta)}{f(\underline{y}; \theta)} = \frac{g(T(\underline{x}; \theta))h(\underline{x})}{g(T(\underline{y}; \theta))h(\underline{y})} = \frac{h(\underline{x})}{h(\underline{y})}$$

□

Theorem 2.7. (On Minimal Sufficiency) Let $\underline{X} \sim f_{\underline{X}}(\underline{x}; \theta)$. Suppose that there exists a statistic, $T = T(\underline{X})$, such that $\forall \underline{x}, \underline{y}$, $\frac{f_{\underline{X}}(\underline{x}; \theta)}{f_{\underline{X}}(\underline{y}; \theta)}$ is independent of θ if and only if $T(\underline{x}) = T(\underline{y})$. Then T is minimal sufficient for θ .

Proof. We first prove minimality (assuming sufficiency) and then go on to prove sufficiency.

Let T^* be any other sufficient statistic for θ . Suppose that $T^*(\underline{x}) = T^*(\underline{y})$. Then, by the lemma, $\frac{f(\underline{x}; \theta)}{f(\underline{y}; \theta)}$ is independent of θ . But by the criterion for T in the statement of the theorem, this means that $T(\underline{x}) = T(\underline{y})$. But this implies that T must be a function of T^* .

Now, we prove sufficiency:

We want to show that $f_{\underline{X}}(\underline{x}|T = t)$ is independent of θ . That is, $f_{\underline{X}}(\underline{x}|T(\underline{x}) = T(\underline{y}))$ is independent of θ .

As before, consider the discrete case:

$$\begin{aligned} P(\underline{X} = \underline{x}|T(\underline{x}) = T(\underline{y})) &= \frac{P(\underline{X} = \underline{x}; \theta)}{P(T(\underline{x}) = T(\underline{y}))} \\ &= \frac{P(\underline{X} = \underline{x}; \theta)}{\sum_{\underline{y}: T(\underline{y}) = T(\underline{x})} P(\underline{X} = \underline{y}; \theta)} \\ &= \frac{1}{\sum_{\underline{y}: T(\underline{y}) = T(\underline{x})} \frac{P(\underline{X} = \underline{y}; \theta)}{P(\underline{X} = \underline{x}; \theta)}} \end{aligned}$$

Which is independent of θ , since we're summing over $\underline{y}$ such that $T(\underline{x}) = T(\underline{y})$. Hence T is sufficient for θ . $\square$

Example: Let X_i be i.i.d. $N(\mu, \sigma^2)$, where μ and σ^2 are unknown. Find a minimal sufficient statistic for (μ, σ^2) .

Solution: We have:

$$\frac{f(\underline{x}; \mu, \sigma^2)}{f(\underline{y}; \mu, \sigma^2)} = e^{\frac{-1}{2\sigma^2}[(\sum X_i^2 - \sum Y_i^2) + 2\mu(\sum X_i - \sum Y_i)]}$$

Which is independent of μ and σ^2 if and only if $\sum_1^n X_i^2 = \sum_1^n Y_i^2$ and $\sum_1^n X_i = \sum_1^n Y_i$.

So therefore, by our theorem, $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is minimal sufficient for (μ, σ^2) .

Example: Let X_i be i.i.d. $U(0, \theta)$ random variables. Find a minimal sufficient statistic for θ .

Solution: Consider:

$$\frac{f(\underline{x}; \theta)}{f(\underline{y}; \theta)} = \frac{\frac{1}{\theta^n} \mathbb{I}_{(0, \theta)}(\max_i x_i) \mathbb{I}_{(0, \theta)}(\min_i x_i)}{\frac{1}{\theta^n} \mathbb{I}_{(0, \theta)}(\max_i y_i) \mathbb{I}_{(0, \theta)}(\min_i y_i)}$$

which is independent of θ if and only if $T(\underline{X}) = \max_i(x_i) = \max_i(y_i) = T(\underline{Y})$. So $T(\underline{X}) = \max_i(x_i)$ is minimal sufficient for θ .

2.2.4 Completeness and a Return to Minimum Variance Unbiasedness

We will see how it is possible to use sufficiency to identify a MVUE, if it exists. We first need the celebrated Rao-Blackwell Theorem.

Theorem 2.8. *Let U be an unbiased estimator of θ . Let T be sufficient for θ .*

Define: $W = W(\underline{X}) = \mathbb{E}(U|T)$. Then:

1. $\mathbb{E}(W) \equiv \theta$ for all $\theta \in \Theta$
2. $\text{Var}(W) \leq \text{Var}(U)$

Note: We define $\mathbb{E}(X|Y)$ as follows by setting $g(y) = \mathbb{E}(X|Y = y)$ then $\mathbb{E}(X|Y) = g(Y)$, which is a random variable, since we are conditioning on a random variable.

Proof. We use the Law of Total Expectation, and the analogue for variances.

1. By the Law of Total Expectation, $\mathbb{E}(W) = \mathbb{E}(\mathbb{E}(U|T)) = \mathbb{E}(U) \equiv \theta$
2. Now, we use that $\text{Var}(Y) = \text{Var}(\mathbb{E}(Y|X)) + \mathbb{E}(\text{Var}(Y|X))$.

Therefore, $\text{Var}(W) = \text{Var}(\mathbb{E}(U|T)) \leq \text{Var}(U)$ for all θ , since $\mathbb{E}(\text{Var}(U|T)) \geq 0$

□

The process of finding $W(T)$ by conditioning an unbiased estimator on a sufficient statistic is called *Rao-Blackwellization*.

Define: A family of distributions, $\{f(x; \theta), \theta \in \Theta\}$ is called *complete* if $\mathbb{E}[g(\underline{X})] = 0 \Rightarrow g(\underline{X}) = 0$ with probability 1. That is, if the only unbiased estimator of 0 is the zero function.

Notes:

1. Completeness is a family property. It is a property not of a single distribution, but of a family of distributions.
2. A statistic, T , is said to be complete if the family of distributions $\{f_T(t; \theta), \theta \in \Theta\}$ is complete.

Theorem 2.9. (*Lehmann - Scheffé*) Let $\{f_T(t; \theta), \theta \in \Theta\}$ be a family of distributions with θ unknown. Let U be an unbiased estimator of θ such that $U = U(T)$ is a function of a complete, sufficient statistic, T . Then U is the unique minimum variance unbiased estimator of θ .

Proof. We prove uniqueness first. We show $U(T)$ is unique amongst all other unbiased estimators (for θ) that are functions of T . Thus, let $U^* = U^*(T)$ be some other unbiased estimator of θ . Then, $\mathbb{E}(U(T)) \stackrel{\theta}{=} \theta$, and $\mathbb{E}(U^*(T)) \stackrel{\theta}{=} \theta$. Therefore, $\mathbb{E}(U(T) - U^*(T)) \stackrel{\theta}{=} 0$. So by completeness, this implies that with probability 1, $U(T) = U^*(T)$. So U is unique amongst all unbiased estimators which are functions of T .

Now we show that it is minimum variance. Let V be any other estimator of θ and suppose that $\text{Var}(V) < \text{Var}(U(T))$. Now, define $W(T) = \mathbb{E}(V|T)$. Then $W(T)$ is unbiased and $\text{Var}(W(T)) \leq \text{Var}(V)$, by Rao-Blackwell. So $\text{Var}(W(T)) < \text{Var}(U(T))$. But by uniqueness of $U(T)$, as a function of T , we have a contradiction. $\square$

Application: We can use the Lehmann - Scheffé Theorem in two different ways to find a unique MVUE.

1. Find an unbiased estimator that is a function of a complete sufficient statistic. (That is, you start with a $U(T)$).
2. Find any unbiased estimator of θ (preferably, as simple as possible). Find a complete sufficient statistic, T , for θ . Define $W(T) = \mathbb{E}(U|T)$

Example: Let $X_1, X_2, \dots, X_n$ be i.i.d. $U(0, \theta)$. Let $Y_n = \max(X_1, \dots, X_n)$. Find the unique minimum variance unbiased estimator.

Solution:

Using method 1:

We saw that $\frac{n+1}{n}Y_n$ is unbiased for θ . Identify T with Y_n , so that $\hat{\theta} = \frac{n+1}{n}Y_n = \hat{\theta}(T)$. We know that $T = Y_n$ is sufficient for θ , and we hope to show that it is complete.

Let g be any function and suppose that $\mathbb{E}(g(Y_n)) \stackrel{\theta}{=} 0$. We must show that $g = 0$ with probability 1. Recall that $f_{Y_n}(y) = \begin{cases} \frac{ny^{n-1}}{\theta^n} & 0 < y < \theta \\ 0 & \text{else, where } \theta > 0 \end{cases}$

So we have:

$$\mathbb{E}[g(Y_n)] = \int_0^\theta g(y) \frac{ny^{n-1}}{\theta^n} dy \stackrel{\theta}{=} 0 \iff \int_0^\theta g(y) y^{n-1} dy \stackrel{\theta}{=} 0$$

So we try the Fundamental Theorem of Calculus,

$$\frac{d}{d\theta} \int_0^\theta g(y) y^{n-1} dy \stackrel{\theta}{=} 0$$

That is, $g(\theta)\theta^{n-1} \stackrel{\theta}{=} 0$, which tells us that $\forall \theta > 0$, we have $g(\theta) \equiv 0$ and therefore $g(y) \equiv 0$ for all $y > 0$.

For $y < 0$, we know that $P(y < 0) = 0$, so $g(Y) = 0$ with probability 1.

Using Method 2:

We show that Y_n is complete and sufficient for θ . We know that $V = 2\bar{X}$ is unbiased for θ . Therefore, $\mathbb{E}(V|Y_n) = 2\mathbb{E}(\bar{X}|Y_n)$ is the unique MVUE for θ .

Its a bit of a pain to work out $2\mathbb{E}(\bar{X}|Y_n)$. However, $\frac{n+1}{n}Y_n$ is also unbiased for θ

and much more convenient, since:

$$\begin{aligned}\mathbb{E}\left(\frac{n+1}{n}Y_n|Y_n\right) &= \frac{n+1}{n}\mathbb{E}(Y_n|Y_n) \\ &= \frac{n+1}{n}Y_n\end{aligned}$$

Which, by the L-S theorem, is the unique MVUE.

2.3 Methods of Estimation

We aim to provide two systematic methods of finding estimators. The idea is that these methods should, in the first place, be “reasonable,” and, in the second place, produce estimators that have “good” properties. Both methods, the method of moments, and the method of maximum likelihood, have good large sample properties. We note that good large sample properties do not ensure good properties for smaller samples.

2.3.1 Method of Moments

Let $X_1, X_2, \dots, X_n$ be i.i.d. $f(\underline{x}; \theta)$ random variables where $\theta = (\theta_1, \dots, \theta_r)$. Suppose that $\mathbb{E}(X^k) = \mu_k = \mu_k(\theta_1, \dots, \theta_r)$ exist for at least $k = 1, 2, \dots, r$.

Definition: The k^{th} sample moment is $m_k = \frac{1}{n} \sum_{i=1}^n X_i^k$. So, for example, $m_1 = \bar{X}$.

Method: Set:

$$\begin{aligned}m_1 &= \mu_1 = g_1(\theta_1, \dots, \theta_r) = \mu_1(\theta_1, \dots, \theta_r) \\ m_2 &= \mu_2 = g_2(\theta_1, \dots, \theta_r) = \mu_2(\theta_1, \dots, \theta_r) \\ &\vdots \\ m_r &= \mu_r = g_r(\theta_1, \dots, \theta_r) = \mu_r(\theta_1, \dots, \theta_r)\end{aligned}$$

Solve, if you can, uniquely, for $\theta_1, \theta_2, \dots, \theta_r$ to get:

$$\begin{aligned}\hat{\theta}_1 &= h_1(m_1, \dots, m_r) \\ &\vdots \\ \hat{\theta}_r &= h_r(m_1, \dots, m_r)\end{aligned}$$

The $\hat{\theta}_k$'s are called the method of moment estimators of the θ_k 's.

General Idea: Set the observed sample moments equal to the corresponding population moments and solve for the θ 's.

Example: Let $X_1, X_2, \dots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ with μ and σ both unknown. Find the method of moment estimators for μ and σ^2 .

Solution:

Proof. We have that $m_1 = \bar{X}$ and $m_2 = \frac{1}{n} \sum_1^n X_i^2$.

Next, $\mathbb{E}(X_1) = \mu(\mu, \sigma^2) = g_1(\mu, \sigma^2)$ and $\mathbb{E}(X_1^2) = \sigma^2 + \mu^2 = g_2(\mu, \sigma^2)$.

Thus, set $m_1 = \bar{X} = \mu$ and $m_2 = \frac{1}{n} \sum_1^n X_i^2 = \sigma^2 + \mu^2$ and solve for μ and σ^2 .

Clearly, $\hat{\mu} = \bar{X}$ and from the second equation, we see that $\frac{1}{n} \sum_1^n X_i^2 = \sigma^2 + \bar{X}^2$, which implies

$$\text{that } \hat{\sigma}^2 = \frac{1}{n} \left[\sum_1^n X_i^2 - n\bar{X}^2 \right] = \frac{1}{n} \left[\sum_1^n (X_i - \bar{X})^2 \right].$$

Since s^2 is known to be unbiased, we can see that $\hat{\sigma}^2$ is *not* unbiased, and therefore, method of moments estimators may not be unbiased. $\square$

The following theorem establishes the consistency of method of moment estimators under suitable conditions.

Theorem 2.10. *In the application of the method of moments, suppose that $h_1, \dots, h_r$ are continuous functions of $(\mu_1, \dots, \mu_r)$. Then the method of moment estimators $\hat{\theta}_i, i = 1, 2, \dots, r$ are consistent for $\theta_i, i = 1, 2, \dots, r$.*

Proof. We have that $\hat{\theta}_i = h_i(m_1, \dots, m_r)$. We must show that $\hat{\theta}_i \xrightarrow{P} \theta_i, \forall i$

Note that, $m_k = \frac{1}{n} \sum_{i=1}^n X_i^k \xrightarrow{P} \mu^k = \mathbb{E}(X^k)$ by the Weak Law of Large Numbers, since μ^k exists.

Since g is continuous, we have that $h_i(m_1, \dots, m_r) \xrightarrow{P} h_i(\mu_1, \dots, \mu_r) = \theta_i$ as $n \rightarrow \infty$. $\square$

Final Comment: Consistency is the only general good property that can be established for method of moments estimators. They are often used because of the simplicity of the method, as initial values applied to better methods of estimation.

2.3.2 Method of Maximum Likelihood

By far, the most commonly used method of estimation is the method of maximum likelihood.

Define: Let $\underline{X} \sim f(\underline{x}; \theta)$, where θ could be vector valued and $\underline{X}$ could be discrete. The *likelihood function* for θ at $\underline{X} = \underline{x}$, written $l(\theta, \underline{x})$ is defined to be $f(\underline{x}, \theta)$ as a function of θ for a given $\underline{x}$.

Note: Even though the likelihood and p.d.f. are the same at a given θ and $\underline{x}$, the p.d.f. is a function of $\underline{x}$ for a given θ , and the likelihood function is the opposite. To find the likelihood, you find the joint p.d.f.

Define: Let $\underline{x}$ be a given sample value. If there exists $\hat{\theta} = \hat{\theta}(\underline{x})$ [one of the values of $\theta \in \Theta$] that maximizes $l(\theta; \underline{x})$ [as a function of θ], then we call $\hat{\theta}$ a *maximum likelihood estimate* of θ and we call $\hat{\theta}(\underline{X})$ a *maximum likelihood estimator* of θ , which we denote *mle* or MLE.

General Idea: Think first of a discrete random variable. The MLE is that of θ that gives the highest probability of observing $\underline{x}$. In the continuous case, $P(\underline{X} = \underline{x}; \theta) = 0$. However, $f(\underline{x}; \theta)d\underline{x} \approx P(x < X \leq x + dx; \theta)$, giving sort of a probabilistic interpretation.

The following simple example illustrates the idea.

Example: Suppose that we have a single Bernoulli observation with unknown p , where $p \in \{\frac{1}{4}, \frac{3}{4}\}$. Find a “good” estimate of p .

Solution: Here $\Theta = \{\frac{1}{4}, \frac{3}{4}\}$ has just 2 values. Suppose that $X = x = 1$.

$P(X = x) = p^x(1 - p)^{1-x}|_{x=1} = p$. Now, we must decide between $p = \frac{1}{4}$ and $p = \frac{3}{4}$. Obviously, choose $p = \frac{3}{4}$ since it is more probable to have observed $\underline{x} = 1$. Indeed, the MLE is $\hat{p} = \frac{3}{4}$

Methods of Finding the MLE:

1. Find the maximum by differentiation. Often it is easier to maximize the log of the likelihood, that is to find the value of θ that maximizes $\log(l)$, which works since $\log$ is a monotone function.
2. Directly search for the maximum (in cases where the maximum falls on the boundary and we have non-differentiability)
3. Use numerical methods.

Using method (1), we need to solve the system of equations,

$$\frac{\partial}{\partial \theta_i} \ln l(\theta_1, \dots, \theta_r; \underline{x}) = 0, i = 1, \dots, r$$

to find the possible extrema. As usual, one must then ascertain if a maximum exists amongst these critical points.

Jargon:

$$\frac{\partial}{\partial \theta_i} \ln l(\theta_1, \dots, \theta_r; \underline{x}) = 0, i = 1, \dots, r$$

is called the *score vector*. The equations are called the *likelihood equations*

Example: Let $X_1, X_2, \dots, X_n$ be i.i.d. $N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. Find the m.l.e. of (μ, σ^2)

Solution: $l((\mu, \sigma^2); \underline{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}} \cdot \frac{1}{(\sigma^2)^{\frac{n}{2}}} e^{\frac{-1}{2\sigma^2} \sum (X_i - \mu)^2}$

So therefore, we have that:

$$\ln(l) = c - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2$$

Which tells us that:

$$\frac{\partial(\ln(l))}{\partial\mu} = 0 \Rightarrow \sum_1^n (X_i - \mu) = 0 \Rightarrow \hat{\mu} = \frac{1}{n} \sum_1^n X_i$$

Next:

$$\frac{\partial(\ln(l))}{\partial\sigma^2} = 0 \Rightarrow \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_1^n (X_i - \bar{X})^2 = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_1^n (X_i - \bar{X})^2$$

We should check that these are maximums but we won't. Hence, after checking to see that $(\bar{X}, \frac{1}{n} \sum_1^n (X_i - \bar{x})^2)$ is a maximum, we find that $(\hat{\mu}, \hat{\sigma}^2)$ is a *mle*.

Exercise: Try finding the *mle* of $X \sim \text{Bern}(p)$ where $\hat{p} = \frac{X}{n}$, $Y \sim \text{Poisson}(\lambda)$ where $\hat{\lambda} = \frac{\sum X_i}{n}$ and $Z \sim \text{Exp}(\beta)$ where $\hat{\beta} = \frac{\sum X_i}{n}$

Warning: Care is required if there is a constraint on the parameter space. Thus, in the above example, with $\sigma^2 = 1$, say. Suppose that you wish to find the m.l.e. under the constraint $\mu \leq 0$.

It is easy to show that the likelihood for the previous example is:

1. Case 1: The m.l.e. is 0, since l is maximized at $\mu = 0$
2. Case 2: The m.l.e. is $\frac{\sum x_i}{n}$

Thus, the m.l.e. is $\min(0, \sum x_i)$

Example: Let $X_1, X_2, \dots, X_n$ be i.i.d. $U(0, \theta)$ random variables. Find the m.l.e. of θ

Solution: Since $f(x; \theta) = \frac{1}{\theta^n} \mathbb{I}_{(0, \theta)}(\max(x_i)) \cdot \mathbb{I}_{(0, \theta)}(\min(x_i))$.

Clearly, by inspection, l is maximized at $\hat{\theta} = \max(x_i)$.

Note that l is not differentiable at $\theta = \max(x_i)$

Example: This example shows that *mle*'s may not be unique: Let $X_1, \dots, X_n$ be i.i.d. $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$ random variables. Find the *mle* at θ .

Solution: $l = \mathbb{I}_{\theta - \frac{1}{2} < \min(x_i) < \max(x_i) < \theta + \frac{1}{2}} \min(x_i) \max(x_i)$

Hence, since the likelihood has a maximum value of 1, any value of θ in that range is an *mle* and so the *mle* is not unique.

Of course, the value of the method of maximum likelihood lies in the properties of the estimators it produces. Before discussing these properties, we should note that the distributional properties of the m.l.e. $\hat{\theta} = \hat{\theta}(\underline{X})$ depend on the distribution of $\underline{X}$. That is, the distribution of the data. Two likelihoods could be proportional and yet their maximum likelihood estimators have entirely different properties.

Example: Suppose a coin is tossed 20 times and we observe 13 heads, where p is unknown. The likelihood is $\binom{20}{13}p^{13}(1-p)^7$. Now, suppose someone else tosses a coin until they get 13 heads and that it took 20 trials for this to happen. Here, the likelihood is $\binom{19}{12}p^{13}(1-p)^7$ (which is Negative Binomial). Note that these two are proportional, and in both cases, we get $\hat{p} = \frac{13}{20}$.

Therefore, in establishing the sampling distribution of $\hat{p} = \frac{X}{N}$, we need to know whether N was fixed and X was random, or whether X was fixed and N was random.

Main Properties of MLE's

1. Asymptotically Unbiased ($\mathbb{E}(\hat{\theta}) \rightarrow \theta$ as $n \rightarrow \infty$)
2. MLE's are consistent for θ when the X_i 's are i.i.d.
3. Invariance - this property holds even if n is small.
4. Asymptotic Normality

2.3.3 Invariance and MLEs

Let $\hat{\theta}$ be the *mle* of θ . Let $\eta = \tau(\theta)$ be any function of θ . Then the *mle* of η , $\hat{\eta}$ say, is $\tau(\hat{\theta})$

Comment: The use of invariance is that you do not have to maximize a likelihood with respect to η . You can simply maximize with respect to θ and find $\hat{\eta} = \tau(\hat{\theta})$

Example: Let $X_1, \dots, X_n$ be i.i.d. $\exp(\beta)$ random variables. Find the *mle* of $Var(X_1)$

Solution: It is easy to see that $Var(X_1) = \beta^2$. We have, almost trivially that $\hat{\beta} = \bar{X}$ So therefore, $\widehat{Var(X_1)} = \bar{X}^2 = \hat{\beta}^2$

(note: this is not s^2)

Example: Suppose that $X_1, \dots, X_n$ are i.i.d. $N(\mu, \sigma^2)$. Find the *mle* of $P(X_1 > 1.8; \mu, \sigma^2)$

Solution: We have that $\hat{\mu} = \bar{X}$ and that $\hat{\sigma}^2 = \frac{1}{n} \sum (X_i - \bar{X})^2$.

Hence, the *mle* is $P(\widehat{X_1} > 1.8) = \int_{1.8}^{\infty} \frac{1}{2\pi} \frac{1}{\hat{\sigma}} e^{-\frac{1}{2}(\frac{x-\hat{\mu}}{\hat{\sigma}})^2} dx$

Theorem 2.11. Let $\hat{\theta}$ be the *mle* of θ . Let $\eta = \tau(\theta)$ be an function of θ , then $\hat{\eta} = \tau(\hat{\theta})$

Proof. We only give the proof for the case that τ is a one-to-one function of θ .

We have that $l^*(\eta; \underline{x}) = f(\tau^{-1}(\eta); \underline{x}) = l(\tau^{-1}(\eta); \underline{x}) = l(\theta; \underline{x})$

Therefore, it is clear that maximizing with respect to η is equivalent to maximizing with respect to θ because η is one-to-one.

So we have:

$$\sup_{\eta} l^*(\eta; \underline{x}) = l^*(\hat{\eta}; \underline{x}) = \sup_{\theta} l(\theta; \underline{x}) = l(\hat{\theta}; \underline{x})$$

Therefore:

$$l^*(\hat{\eta}; x) = l(\tau^{-1}(\tau(\hat{\theta}); x))$$

And so we conclude our result: $\hat{\eta} = \tau(\hat{\theta})$ □

Asymptotic Normality of MLE's

Theorem 2.12. Let $X_1, X_2, \dots, X_n$ be i.i.d. random variables with pdf $f(x; \theta)$ where θ is scalar valued. Define $I(\theta) = \mathbb{E}[(\frac{\partial}{\partial \theta} \ln f_X(x; \theta))^2]$.

Then $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{D}} Z \sim N(0, \frac{1}{I(\theta)})$

Comment: Since it can be shown that $\mathbb{E}(\hat{\theta}) \rightarrow \theta$, it follows by the C-R inequality that asymptotically, MLE's have minimum variance amongst asymptotically unbiased estimators. We say that for i.i.d. X_i 's, *mle*'s are asymptotically most efficient. Furthermore, we call $I(\theta)$ the *Fisher information*.

Proof. First, we need the following results:

1. $\mathbb{E}_{\theta_0}[\frac{\partial}{\partial \theta} \ln l(\theta; \underline{x})|_{\theta=\theta_0}] = 0$ where θ_0 is the “true” θ that defines the distribution of the X_i 's
2. $Var_{\theta_0}[\frac{\partial}{\partial \theta} \ln l(\theta; \underline{x})|_{\theta=\theta_0}] = \mathbb{E}_{\theta_0}[(\frac{\partial}{\partial \theta} \ln l(\theta; \underline{x}))^2|_{\theta=\theta_0}]$ which follows immediately from (1)

3. $\mathbb{E}_{\theta_0}[\frac{\partial}{\partial \theta} \ln l(\theta; \underline{x})^2 |_{\theta=\theta_0}] = -\mathbb{E}_{\theta_0}[\frac{\partial^2}{\partial \theta^2} \ln l(\theta; \underline{x}) |_{\theta=\theta_0}]$
4. $\frac{-1}{n} \frac{\partial^2}{\partial \theta^2} \ln l(\theta; \underline{x}) \longrightarrow -\mathbb{E}_{\theta_0}[\frac{\partial^2}{\partial \theta^2} \ln l(\theta; \underline{x}) |_{\theta=\theta_0}]$ in probability (by the WLLN) for all θ and therefore $\theta = \theta_0$

Now, the idea of the proof is the use a 1st order Taylor expansion for $\frac{\partial}{\partial \theta} \ln l(\theta; \underline{x})$

Let θ_0 denote the true parameter value of θ . Let $\hat{\theta}$ denote the *mle* and suppose that n is large so that $\hat{\theta}$ is close to θ_0 (because $\hat{\theta}$ is consistent).

We have:

$$0 \stackrel{mle}{=} \frac{\partial}{\partial \theta} \ln l(\theta; \underline{x})|_{\theta=\theta_0} \approx \frac{\partial}{\partial \theta} \ln l(\theta; \underline{x})|_{\theta=\theta_0} + (\hat{\theta} - \theta_0) \frac{\partial^2}{\partial \theta^2} \ln l(\theta; \underline{x})|_{\theta=\theta_0}$$

That is:

$$(\hat{\theta} - \theta_0) \frac{\partial^2}{\partial \theta^2} \ln l(\theta; \underline{x})|_{\theta=\theta_0} = -\frac{\partial}{\partial \theta} \ln l(\theta; \underline{x})|_{\theta=\theta_0}$$

So subtracting $\mathbb{E}_{\theta_0}[\frac{\partial}{\partial \theta} \ln l(\theta; \underline{x})|_{\theta=\theta_0}] = 0$ from both sides and dividing by $\sqrt{\text{Var}_{\theta_0}[\frac{\partial}{\partial \theta} \ln l(\theta; \underline{x})|_{\theta=\theta_0}]}$

The right hand side becomes:

$$\frac{\sum_{i=1}^n \frac{\partial}{\partial \theta} \ln l(\theta; x_i)|_{\theta=\theta_0} - \mathbb{E}_{\theta_0}[\frac{\partial}{\partial \theta} \ln l(\theta; \underline{x})|_{\theta=\theta_0}]}{\sqrt{\text{Var}_{\theta_0}[\frac{\partial}{\partial \theta} \ln l(\theta; \underline{x})|_{\theta=\theta_0}]}}$$

Since $\frac{\partial}{\partial \theta} \ln l(\theta; x_i)|_{\theta=\theta_0}$ are i.i.d. random variables with presumed finite variances and means. The central limit theorem implies that the left hand side converges in distribution to $Z \sim N(0, 1)$

So invoking (4) and dividing by $\frac{1}{n}$ to get $I(\theta)$, we conclude that $\frac{\sqrt{n}(\hat{\theta}-\theta_0)}{\frac{1}{I(\theta)}} \sim N(0, 1)$ and so we conclude our result. $\square$

Comment: We have that:

$$\frac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{I^{-1}(\theta)}} = \frac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{I^{-1}(\theta)}} = \frac{\sqrt{n}(\hat{\theta} - \theta)}{\left(\sqrt{\frac{-\mathbb{E}[\frac{\partial^2}{\partial \theta^2} \ln f(\underline{x}; \theta)]}{n}} \right)^{-1}}$$

Now observe that:

$$\frac{\frac{\partial^2}{\partial \theta^2} \ln f(\underline{x}; \theta)}{n} = \frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \ln f(x_i; \theta) \xrightarrow{WLLN} I(\theta)$$

Definition: We call $-\frac{\partial^2}{\partial \theta^2} \ln f(x_1; \theta)$ the *observed information on θ* , based on the single observation of X_1 . Call this $\mathcal{I}(\theta)$, but note that this definition is slightly useless, because we cannot replace $I(\theta)$ with $\mathcal{I}(\theta)$

The conclusion is that $\frac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{\frac{\mathcal{I}_n(\theta)}{n}}} \xrightarrow{\mathcal{D}} Z \sim N(0, 1)$ where $\mathcal{I}_n(\theta)$ is the observed information for n observations.

2.4 Interval Estimation

Up until now, we have concentrated on so-called point estimators, which will provide a point in $\mathbb{R}^k$ (where k is the dimension of θ) once we have collected the data. We have sought estimators that have good “average” properties over repeated uses. We can say nothing about their performances on a specific occasion once the data has been collected.

Instead of seeking a single “guess” at an estimator of θ , we now seek a random interval that will, with high (pre-determined) probability, capture the true θ .

Specifically, we seek endpoints, L and R such that $P(L < \theta < R) = 1 - \alpha$ where α is pre-determined and selected to be small.

Definition: An interval $L(\underline{X}), R(\underline{X})$ is called a $100(1 - \alpha)\%$ *confidence interval for θ* if $P(L(\underline{X}) < \theta < R(\underline{X})) = 1 - \alpha$ for all $\theta \in \Theta$ and provided L and R depend only on $\underline{X}$ and possibly on *known* parameters.

Notes:

1. The interval $(L(\underline{X}), R(\underline{X}))$ is random, before we collect the data. Hence, this interval could be called a probability interval. In this case, probability is well defined. However, the name confidence interval is still inappropriately used for this random interval.
2. $1 - \alpha$ is called the *confidence level*

3. Once the data $\underline{X} = \underline{x}$ are observed, the interval $(L(\underline{x}), R(\underline{x}))$ is fixed. Therefore, the only meaningful statement about θ and thus the interval is $P(L(\underline{x}) < \theta < R(\underline{x}))$ is 1 if the fixed θ falls in the interval or 0 if not.

The interval $(L(\underline{x}), R(\underline{x}))$ is also called a $100(1 - \alpha)\%$ *confidence interval for θ*

4. Even though one refers to $100(1 - \alpha)\%$ confidence, the term confidence has no mathematical definition. The “confidence” comes from the high probability that θ is in this interval.
5. Never talk about finding a confidence interval for $\bar{X}$ or s^2 or any function of the data. Confidence intervals are for *unknown* parameters. We don’t need to make inference about $\bar{X}$, as we can observe it.

2.4.1 Frequentist Interpretations

First, we give the weak interpretation.

If I have an unknown θ and if I repeatedly collect data sets, each time a computing a confidence interval (95% say) over a large number of such data sets, then roughly 95% of them will contain θ

The strong interpretation is as follows:

First, we have θ_1 , collect data and find a confidence interval for θ_1 . Then, we have θ_2 , collect data and find a confidence interval for θ_2 , and so on. Note that the θ_i ’s are different values of θ in the same experimental set-up. The strong interpretation is that 95% of these intervals will capture their respective θ_i ’s

The reason that the strong interpretation is better is that we have that $P(L(\underline{X}) < \theta < R(\underline{x})) = 1 - \alpha$ for all $\theta \in \Theta$.

Now, define random variables:
$$Y_i = \begin{cases} 1 & \theta_i \in (L_i(\underline{X}), R_i(\underline{X})) \\ 0 & \text{else} \end{cases}$$

The Y_i ’s are i.i.d. Bernoulli random variables since the experiments are assured independent and $P(Y_i = 1) = P(L(\underline{X}) < \theta_i < R(\underline{x})) = 1 - \alpha$ for all i .

Therefore, $\frac{1}{n} \sum_{i=1}^n Y_i \xrightarrow{P} P(Y_i = 1) = \mathbb{E}(Y_i) = 1 - \alpha$ where the l.h.s. is the proportion of intervals that capture θ .

2.4.2 The Actual Computation

The main method that we'll use is that of finding a pivotal quantity $g(\underline{x}; \theta)$ such that the distribution of $g(\underline{x}; \theta)$ can be found but it is independent of θ (the *distribution* is independent of θ)

With such a pivotal quantity, we can write $P(a_{\gamma} < g(\underline{x}; \theta) < b_{\gamma'}) = 1 - \alpha$ where a_{γ} and $b_{\gamma'}$ are selected to have this property. Then, hopefully, one can invert the inequalities to obtain θ in the middle such that $P(L(\underline{x}; a_{\gamma}) < \theta < R(\underline{x}; b_{\gamma'})) = 1 - \alpha$.

2.4.3 The Two-Sample Setup

Sometimes we have two populations and are interested in comparing their respective parameters (perhaps $\mu_1 - \mu_2$ or $\frac{\sigma_1^2}{\sigma_2^2}$). Thus we take $\underline{X}$ from population 1 and $\underline{Y}$ from population 2, and derive a confidence interval for the comparison.

Example: Let $X_1, X_2, \dots, X_{n_1}$ be i.i.d. $N(\mu_1, \sigma^2)$ random variables. Let $Y_1, Y_2, \dots, Y_{n_2}$ be i.i.d. $N(\mu_2, \sigma^2)$ random variables. Let $\underline{X} \perp \underline{Y}$. Show that a $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by $\bar{X} - \bar{Y} \pm t_{n_1+n_2-2}(\frac{\alpha}{2}) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ where $s_p = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$, the pooled sample variance, and s_1 and s_2 are the sample variances of the X's and Y's respectively.

Solution: We seek L and R such that $P(L < \mu_1 - \mu_2 < R) = 1 - \alpha$

That is:

$$P\left(\frac{\bar{X} - \bar{Y} - R}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < \frac{\bar{X} - \bar{Y} - L}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right) = 1 - \alpha$$

Claim:

$$\underbrace{\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}_A \sim t_{n_1+n_2-2}$$

Proof.

$$A = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} \bigg/ \sqrt{\frac{\frac{(n_1+n_2-2)s_p^2}{\sigma^2}}{n_1 + n_2 - 2}}$$

The numerator and denominator of this giant monster are independent since $\bar{X} - \bar{Y}$ and s^2 are independent. Now, it remains to show that the numerator of the denominator is $\chi_{n_1+n_2-2}^2$. If so, we're done.

Recall that: $s_p = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

We see that the numerator of the denominator is equal to:

$$\left(\underbrace{\frac{(n_1 - 1)s_1^2}{\sigma^2}}_{\chi_{n_1-1}^2} + \underbrace{\frac{(n_2 - 1)s_2^2}{\sigma^2}}_{\chi_{n_2-1}^2} \right) / (n_1 + n_2 - 2)$$

Now, if we multiply by $(n_1 + n_2 - 2)$, we get a $\chi_{n_1+n_2-2}^2$ □

Important Special Cases:

1. Let $X_1, X_2, \dots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ with μ and σ^2 unknown. Find a $100(1 - \alpha)\%$ confidence interval for μ , regarding σ^2 as a “nuisance” parameter.

Solution: We seek an L and R such that $P(L < \mu < R) = 1 - \alpha$, that is, such that

$$P\left(\frac{\bar{X} - R}{\frac{s}{\sqrt{n}}} < \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} < \frac{\bar{X} - L}{\frac{s}{\sqrt{n}}}\right) = 1 - \alpha$$

But we know that $\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$.

Suppose we agree to use a confidence interval that is symmetric about $\bar{X}$

We have, from the symmetry of the t-distribution about 0, that:

$$P(-t_{n-1}(\frac{\alpha}{2}) < T_{n-1} < t_{n-1}(\frac{\alpha}{2})) = 1 - \alpha$$

Therefore set two inequalities equal to each other, and solve for L and R .

We get $L = \bar{X} - t_{n-1}(\frac{\alpha}{2})\frac{s}{\sqrt{n}}$ and $R = \bar{X} + t_{n-1}(\frac{\alpha}{2})\frac{s}{\sqrt{n}}$

2. Let $X_1, X_2, \dots, X_n$ be i.i.d. with mean μ and variance σ^2 both unknown. Find a $100(1 - \alpha)\%$ confidence interval for μ . Let n be large (≥ 30)

Solution: We seek an L and R such that $P(L < \mu < R) = 1 - \alpha$, that is, such that

$$P\left(\frac{\bar{X} - R}{\frac{s}{\sqrt{n}}} < \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} < \frac{\bar{X} - L}{\frac{s}{\sqrt{n}}}\right) = 1 - \alpha$$

But for large n , by the CLT, $\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \rightarrow N(0, 1)$

We have by definition that $P(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}) = 1 - \alpha$

So set:

$$\begin{aligned} -z_{\frac{\alpha}{2}} &= \left(\frac{\bar{X} - R}{\frac{s}{\sqrt{n}}}\right) \\ z_{\frac{\alpha}{2}} &= \left(\frac{\bar{X} - L}{\frac{s}{\sqrt{n}}}\right) \end{aligned}$$

and solve for L and R

3. Let $X_1, \dots, X_n$ be i.i.d. $Bern(p)$ random variables with p unknown. Find a $100(1 - \alpha)\%$ confidence interval for p . Let n be large.

Solution: Let $\hat{p} = \frac{X}{n} = \frac{\sum X_i}{n}$ Then the confidence interval is given by $\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Proof is on one of the assignments.

4. Let $X_1, \dots, X_n$ be i.i.d. $exp(\beta)$ random variables. Find an exact $100(1 - \alpha)\%$ confidence interval for β

Comment: Some confidence intervals are *exact* in the sense that the coverage probability $(1 - \alpha)$ is exact else, if we have used a large sample and the CLT, they are called *approximate* since the probability only holds asymptotically.

Solution: We seek an L and R such that $P(L < \beta < R) = 1 - \alpha$. We need a pivotal quantity.

We know that the m.g.f. of X_1 is $M_{X_1}(t) = \frac{1}{1 - \beta t}$

Furthermore, $M_{\sum X_i}(t) = \frac{1}{(1-\beta t)^n}$ so we get $M_{\frac{2\sum X_i}{\beta}}(t) = M_{\sum X_i}\left(\frac{2t}{\beta}\right) = \frac{1}{(1-2t)^n}$

But this distribution is independent of β so the pivotal quantity is $\frac{2\sum X_i}{\beta}$ whose distribution is χ_{2n}^2

5. Now suppose that the X 's and the Y 's do not have necessarily normal distributions. Suppose that n_1 and n_2 are large, then a $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by $\bar{X} - \bar{Y} \pm z_{\frac{\alpha}{2}} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
6. Now suppose that σ_1^2 is not necessarily equal to σ_2^2 but n_1 and n_2 are large. Then a $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is $\bar{X} - \bar{Y} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
7. Let $X_1, \dots, X_{n_1}$ be i.i.d. $Bern(p_1)$ and let $Y_1, \dots, Y_{n_2}$ be i.i.d. $Bern(p_2)$ random variables. Let $\underline{X} \perp \underline{Y}$ and let n_1 and n_2 be large. Then a $100(1 - \alpha)\%$ confidence interval for $p_1 - p_2$ is given by $\hat{p}_1 - \hat{p}_2 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

2.4.4 Sample Size Calculation

Frequently in designing a study, one is required to specify how many observations should be taken to ensure a confidence interval with a given coverage probability $1 - \alpha$ and a given precision (i.e. width)

To determine the sample size that will need to be taken to achieve a specified coverage probability for a confidence interval with a specified width, we would need to set the width, $w = 2 \cdot z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ and solve for n . That is, $n = \frac{4z_{\frac{\alpha}{2}}^2 \hat{p}(1-\hat{p})}{w^2}$ but $\hat{p}$ depends on the data yet to be collected. So instead, we must hope that we have some idea of a reasonable $p = p_0$ from a previous experiment. Or, alternatively, replace $\hat{p}$ with a p_0 that gives you the largest (that is, most conservative) sample size you would need. It is easily seen that for $0 \leq x \leq 1$, $f(x) = x(1 - x)$ has a maximum at $x = \frac{1}{2}$. Hence, set $\hat{p} = \frac{1}{2}$ to get $n = \frac{z_{\frac{\alpha}{2}}^2}{w^2}$

Example: Suppose we shall observe observations that will be normally distributed with known variance σ^2 . What sample size should we take to ensure that we will get a confidence interval for μ of width w and coverage probability $1 - \alpha$?

Solution: We know that a confidence interval for μ under this setup is given by: $\bar{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

so therefore $w = R - L = \frac{2z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}$ so solving, we get $n = \frac{4z_{\frac{\alpha}{2}}^2\sigma^2}{w^2}$

Note: If σ^2 is unknown, we cannot use s^2 instead of σ^2 , since s^2 depends on the data not yet collected. We could, though, develop a conservative estimate of σ^2 from a previous study.

Example: Now suppose that $X_1, X_2, \dots, X_n$ are i.i.d. $Bern(p)$. Find the sample size necessary to ensure a $100(1 - \alpha)\%$ confidence interval that has width, w .

Solution: The confidence interval is given by $\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$ so set $n = \frac{4z_{\frac{\alpha}{2}}^2 p(1-p)}{w^2}$ and since p is unknown, a conservative n will be given by inserting the maximum of $p(1-p)$ for $0 < p < 1$, that is $p = \frac{1}{2}$

2.4.5 Opinion Polls:

In an opinion poll, the parameter of interest is often the “the proportion of people that are in favor of some proposition.”

Example: We interview 1000 Canadians and ascertain that 480 are in favor of fluffy kittens. First we need to write down a point estimate of the proportion of Canadians in favor of fluffy kittens. Then, find a 95% confidence interval for this proportion.

Solution: Let p be the proportion of Canadians in favor of fluffy kittens. Then, $\hat{p} = \frac{480}{1000} = 0.48$ (We know of course that $\hat{p}(X) = \frac{X}{n}$, the random variable, has good properties.) For the second question, the random confidence interval, $\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ has approximate coverage probability of 0.95. Therefore, the 95% confidence interval that we obtain on this occasion is given by $.48 \pm z_{0.025} \sqrt{\frac{0.48(1-0.48)}{1000}}$

Example: In the media, most often the result of the poll is reported as follows: “The poll based on 1000 subjects found that 48% were in favor of capital punishment. This poll is accurate within 3 percentage points 19 times out of 20.”

1. Results obtained from a particular observation have no probability attached to them
2. The “19 times out of 20” is meant to refer to a coverage probability of 0.95
3. “is accurate within 3 percentage points” is meant to be interpreted as follows: The maximum width that can be obtained for a 95% confidence interval for p , based on $n = 1000$ is $2 \cdot 0.03 = 0.06$. Since $\hat{p}$ will be in the center of each confidence interval the

maximum error that can be incurred by using $\hat{p}$ to estimate p is 0.03 (with probability 0.95)

Interpretation: The procedure (that is, the random interval) that led to our current fixed interval will capture p 95% of the time when the sample size is large. We can say nothing about the current interval other than it catches p with probability 0 or 1, and that it was calculated from a random interval that catches p with probability 0.95

2.5 Hypothesis (or Significance) Testing

Instead of attempting to estimate an unknown parameter, θ (either by a point estimator or through an interval) we instead might wish to decide between two choices for θ .

For example, in tossing a coin, is:

1. $p = \frac{1}{2}$ or $p = \frac{3}{4}$
2. $p < \frac{1}{2}$ or $p \geq \frac{1}{2}$
3. $p = \frac{1}{4}$ or $p > \frac{1}{4}$

We must use our data to decide between our two mutually exclusive possibilities for θ . When using the data to make a decision which θ to choose, we immediately see that there is no guarantee that we will make the right choice. Having decided before the experiment how we will choose between $p = \frac{1}{2}$ and $p = \frac{3}{4}$ there is a chance that after the experiment we will choose the wrong one.

The idea is to construct rules of choice so that if they are applied repeatedly, they will lead to decision errors with small probability.

Formalities:

Let Θ be a parameter space and suppose that $\Theta = \Theta_0 \cup \Theta_1$ where $\Theta_0 \cap \Theta_1 = \emptyset$. We call the hypothesis $\theta \in \Theta_0$ the *null hypothesis* and denote it by $H_0 : \theta \in \Theta_0$. The hypothesis that $\theta \in \Theta_1$ is called the *alternative hypothesis* and is denoted by $H_a : \theta \in \Theta_1$ (or often H_1)

A hypothesis test consists of two hypotheses, $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$. We wish to decide between H_0 and H_1 .

We shall divide the sample space of $\underline{X}$ into two regions, A and R , such that $S = A \sqcup R$ and such that if $\underline{X} \in A$ we will “accept” H_0 , and if $\underline{X} \in R$ we will reject H_0 and accept H_a . We also call R a *critical region*.

In practice, we shall use some statistic or function of the data to make our decision. Thus, we will use $T = T(\underline{X})$, a *test statistic* and divide its range into an acceptance region and rejection region. Thus if $T \in A_T$ we will “accept” the null hypothesis, while if $T \in R_T$ we will reject the null hypothesis.

Now we wish to minimize the probability that we will make a mistake. We know that no matter how A_T and R_T are defined, there will inevitably be the possibility that when a decision is made of either accepting or rejecting H_0 , we may in fact choose incorrectly.

If we reject when H_0 is true, we are said to have committed a *Type 1 error*. If we accept when H_a is true, we are said to have committed a *Type 2 error*.

With the realization that either one of these errors could be committed, the aim is to choose R_T and hence A_T such that the probability of making a Type 1 error is tolerably small and the same for a Type 2 error.

Definition: We call $\sup_{\theta \in \Theta_0} P(T \in R_T | \theta) = \alpha$ the *size or significance level* of the test. Note that the notation does not mean conditional probability, but just that the probability depends on the parameter θ .

In the coin tossing example (of choosing between $P(\text{heads}) = \frac{1}{2}$ or $\frac{3}{4}$), if the rejection rule is: Reject if $T \geq 8$, then $\alpha = P(T \geq 8 | p = \frac{1}{2})$ where T is the number of heads out of 10.

Definition: For any given $\theta \in \Theta_1$, denote $P(T \in A | \theta \in \Theta_1)$ by β_θ . This is the probability of a Type 2 error under $\theta \in \Theta_1$

For example, $P(T < 8 | p = \frac{3}{4})$ gives the probability of a Type 2 error in the coin problem with $p = \frac{3}{4}$

Definition: $1 - \beta_\theta = P(T \in R | \theta \in \Theta_1)$ is called the *power of the test at $\theta \in \Theta_1$*

We would like both α and β to be small. Unfortunately, as α decreases, β increases.

Definition: We call the function of θ , $\pi(\theta) \equiv P(T \in R | \theta \in \Theta)$ the *power function* of the test. Note that the power of the test refers only to $\theta \in \Theta_1$ while the power function refers to all θ .

In practice, we proceed as follows: We decide which of the two errors is considered to be more serious. With this in mind, we then define H_0 and H_a so that the Type 1 error is the more serious type of error.

We then define R so that the significance level is small (usually $\alpha \leq 0.05$)

Now, what do we do with the Type 2 error probability? It might be large when R is chosen so that $\alpha = 0.05$, say. We can control the power of a test by increasing the number of observations, as we will see shortly, but sometimes we cannot do this.

Summary: First, construct R so that α is small. If possible - try to ensure that the power is reasonable (usually ≥ 0.80) by increasing n or by other means. Since α is generally under control, and β is not, it will frequently happen that β is large, and correspondingly, the power is small. For this reason, if $T \in A$ we do not accept H_0 , but instead we chicken out and declare that “there is no evidence to reject H_0 at the $100\alpha\%$ level of significance.” However, if β is small (≤ 0.05) and $T \in A$, then we can say “We accept H_0 , with a Type 2 error probability of β ”

Jargon: If we have set up a rejection region of size α , whenever $T \in R$, we declare: “There is evidence to reject H_0 and accept H_a at the $100\alpha\%$ significance level.” We refer to significance of $\alpha \leq 0.05$ by convention only.

Important: On any particular occasion, we have observed the data, if $T_{obs} \in R$, it is incorrect to say that we can reject H_0 and accept H_a because there will be only a probability of α of making a Type 1 error. Either we have or we have not made a Type 1 error. It is the procedure (the rejection region) that has the probability of α .

Notes:

1. H_0 and H_a must be decided on in advance of seeing the data.

For example: We are not sure whether to test $H_0 : \mu = \mu_0 = 1.2$ vs $H_a : \mu > \mu_0 = 1.2$ or to test $H_0 : \mu = \mu_0 = 1.2$ vs $H_a : \mu < \mu_0 = 1.2$. We then observe $\bar{X}$ is 5.7, and conclude that you should test $H_0 : \mu = \mu_0 = 1.2$ vs $H_a : \mu > \mu_0 = 1.2$ since to test $\mu < 1.2$ would be insane.

This is an incorrect argument, and is simply wrong. Here, we have used the data to determine H_a .

Briefly: If you do this, then you should find your significance level as $P_{\mu=1.2}(T \in R|\bar{X} = 5.7) >> \alpha$. So doing this will make α be too big.

2. α has to be determined before seeing the data. You are stuck with it no matter where T falls. Having chosen α , R is then determined, again before seeing the data.
3. Never test hypotheses about statistics (e.g. $\bar{X}$, s^2 , etc.) We don't need to test about these. We can see their values from the data. Tests are about population parameters.
4. $\alpha \neq P(H_0)$ or $P(H_a)$. In “frequentist” hypothesis testing, we never find probabilities of hypotheses. Rather, we find probabilities of the data given the hypothesis (How consistent is my data with the hypothesis we're testing?)

Example: Let $X_1, X_2, \dots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ with σ^2 unknown. Find an α -level test of $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$ where μ_0 is specified.

(This is a so-called *one-sided* test of the mean.)

Solution: Define $T = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$ where s is the sample standard deviation. Define the rejection rule as follows: Reject H_0 and accept H_a if $T > t_{n-1}(\alpha)$

We claim that $P(T \in R | \mu = \mu_0) = \alpha$. That is, with this R , the probability of a Type 1 error is α , as required.

But we've seen before that $\frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$ under H_0

Comment: This particular rejection rule (or test) will be seen to have “good” properties. It is called the *one-sided t-test*.

Note that if H_a were $H_a : \mu < \mu_0$ then we would reject H_0 if $T = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} < -t_{n-1}(\alpha)$

Finally, if $H_a : \mu \neq \mu_0$, then we would reject for either small or large values of T . Choose a rejection rule: Reject H_0 if $|T| > t_{n-1}(\frac{\alpha}{2})$

Word Problem: A certain company makes high precision parts that they claim have a mean diameter of 1.50 mm, with an unknown standard deviation. It is assumed that the diameters are roughly normally distributed. A new process is introduced and it is suspected that the mean diameter may have increased. To test this hypothesis, 10 parts are manufactured and it is found that $\bar{X}_{obs} = 1.51$ mm and $s = 0.01$.

At the 5% level, test a suitable hypothesis.

Solution: Let μ be the mean diameter. We have: $H_0 : \mu = 1.5$ vs $H_a : \mu > 1.5$. We use $T = \frac{\bar{X}-1.5}{\frac{s}{\sqrt{n}}}$

Our rejection rule is: Reject if $T > t_{10-1}(0.05) = t_9(0.05) = 1.833$

We therefore calculate: $T_{obs} = \frac{1.51-1.50}{\frac{0.1}{\sqrt{10}}} = \sqrt{10} > 1.833$

Conclusion: Since $T_{obs} \in R$, we reject H_0 and accept H_a that $\mu > 1.5$, at the 5% significance level. The conclusion is that based on data collected, there is strong evidence that the mean is greater than 1.5, and the procedure used has low probability of being wrong.

Example: Now, change the above problem as follows: Let $s = 0.1$. Now we find $T_{obs} = \frac{1}{\sqrt{10}} < 1.833$. So $T \notin R$ and there is no evidence to reject H_0 that $\mu = 1.5$ at the 5% level of significance.

2.5.1 The Construction of Optimal Tests Under Certain Circumstances

The aim is to derive a procedure that will, under certain circumstances, will produce tests with “good” properties. Even if we cannot show that such procedures are optimal in all circumstances, nevertheless, they suggest how we may construct a test.

First, some terminology:

Definition: If Θ_0 consists of a single hypothesized parameter, then we talk of a *simple* null and alternative hypothesis. Otherwise, the hypotheses are termed *composite*

Examples:

1. $H_0 : \mu = \mu_0$ vs $H_a : \mu = \mu_1$ is a simple vs a simple if σ^2 is known
2. $H_0 : \mu \geq \mu_0$ vs $H_a : \mu < \mu_0$ is a composite vs a composite whether or not σ^2 is known.
3. $H_0 : \mu = \mu_0$ vs $H_a : \mu = \mu_1$ is a composite vs a composite if σ^2 is unknown. (Since $\Theta_0 = \{(\mu, \sigma^2) | \mu = \mu_0, \sigma^2 > 0\}$ and the same for Θ_1)

Let R be a rejection region with associated power $\pi(\theta)$ for each $\theta \in \Theta_1$.

Definition: The test given by R is a *most powerful α -level test in testing $H_0 : \theta = \theta_0$ vs $H_a : \theta = \theta_1$* , if $P(R|\theta_0) = \alpha$ and if R^* is any other test, with power $\pi^*(\theta)$ and $P(R^*|\theta_0) = \alpha$ then $\pi^*(\theta_1) \leq \pi(\theta_1)$

Neyman-Pearson Lemma:

The following celebrated lemma shows how to construct a most-powerful α -level test of a simple vs simple hypothesis.

Lemma 2.13. (Neyman-Pearson) Consider the hypothesis $H_0 : \theta = \theta_0$ vs $H_a : \theta = \theta_1$. Let the rejection region, R , be defined as $R = \{\underline{x} : \frac{f(\underline{x}|\theta_0)}{f(\underline{x}|\theta_1)} \leq c_\alpha\}$ for some constant c_α such that $P(\frac{f(\underline{x}|\theta_0)}{f(\underline{x}|\theta_1)} \leq c_\alpha | \theta_0) = \alpha$. Then R is the most powerful α -level test of H_0 vs H_a .

Note: There are two components to the test: The form of the rejection rule: Reject H_0 in favor of H_1 if the ratio is less than or equal to some constant. Second, choose the constant c_α so that with this rule, the probability of a Type 1 error is the given α

Proof. (Continuous Case) Let $\pi(\theta_1)$ be the power of the test defined by R (at θ_1). Let $\pi^* = \pi^*(\theta_1)$ be the power of any other α -level test defined by the rejection region R^* . We must show that $\pi(\theta_1) \geq \pi^*(\theta_1)$

First:

$$\alpha = P(R|\theta_0) = \int_R f(\underline{x}|\theta_0)d\underline{x} = \int_{R \cap R^*} f(\underline{x}|\theta_0)d\underline{x} + \int_{R \cap R^{*c}} f(\underline{x}|\theta_0)d\underline{x}$$

We also have (since R^* is also α level):

$$\alpha = \int_{R^*} f(\underline{x}|\theta_0)d\underline{x} = \int_{R \cap R^*} f(\underline{x}|\theta_0)d\underline{x} + \int_{R^* \cap R^c} f(\underline{x}|\theta_0)d\underline{x}$$

Therefore:

$$\int_{R \cap R^{*c}} f(\underline{x}|\theta_0)d\underline{x} = \int_{R^* \cap R^c} f(\underline{x}|\theta_0)d\underline{x}$$

Now:

$$\pi(\theta_1) = P(R|\theta_1) = \int_R f(\underline{x}|\theta_1)d\underline{x} = \int_{R \cap R^*} f(\underline{x}|\theta_1)d\underline{x} + \int_{R \cap R^{*c}} f(\underline{x}|\theta_1)d\underline{x}$$

So we get:

$$\begin{aligned}
\int_{R \cap R^{*c}} f(\underline{x}; \theta_1) d\underline{x} &\geq \frac{1}{c_\alpha} \int_{R \cap R^{*c}} f(\underline{x}; \theta_0) d\underline{x} \\
&= \frac{1}{c_\alpha} \int_{R^* \cap R^c} f(\underline{x}; \theta_0) d\underline{x} \\
&\geq \frac{1}{c_\alpha} \int_{R^c \cap R^*} c_\alpha f(\underline{x}; \theta_1) d\underline{x} \\
&= \int_{R^* \cap R^c} f(\underline{x}; \theta_1) d\underline{x}
\end{aligned}$$

So therefore:

$$\begin{aligned}
\pi(\theta_1) &= \int_{R \cap R^*} f(\underline{x}|\theta_1) d\underline{x} + \int_{R \cap R^{*c}} f(\underline{x}|\theta_1) d\underline{x} \\
&\geq \int_{R \cap R^*} f(\underline{x}|\theta_1) d\underline{x} + \int_{R^* \cap R^c} f(\underline{x}|\theta_1) d\underline{x} \\
&= \int_{R^*} f(\underline{x}; \theta_1) d\underline{x} \\
&= \pi^*(\theta)
\end{aligned}$$

□

Example of the Use of the N-P Lemma Let $X_1, \dots, X_n$ be i.i.d. $N(\mu, \sigma^2)$, with σ^2 known. Find the most powerful test of $H_0 : \mu = \mu_0$ vs $H_a : \mu = \mu_1$ where $\mu_1 > \mu_0$.

Solution: Since we have a simple vs simple, the MP α -level test says to reject if we have c_α such that $\frac{f(\underline{x}; \mu_0)}{f(\underline{x}; \mu_1)} \leq c_\alpha$

That is (after some algebraic manipulation), we should reject if $\bar{X} \geq c^*$. The question is, what is this mysterious constant?

We need:

$$\alpha = P(\bar{X} \geq c_\alpha | \mu_0) = P\left(\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \geq \frac{c_\alpha - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right)$$

But, under H_0 , we have that $\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$. So we can set $\frac{c_\alpha - \mu_0}{\frac{\sigma}{\sqrt{n}}} = z_\alpha$ and so we get

$$c_\alpha = z_\alpha \cdot \frac{\sigma}{\sqrt{n}} + \mu_0$$

While this example is nice to work out, we note that simple vs simple hypotheses are not common in practice, so we need at least to consider simple vs composite hypotheses.

Definition: In testing $H_0 : \theta = \theta_0 = \Theta_0$ vs $H_a : \theta \in \Theta_1$ a test is said to be a *uniformly most powerful (UMP) α -level test* if for any other α -level test, with rejection region R^* and power $\pi^*(\theta)$ for $\theta \in \Theta_1$, we have $\pi(\theta) \geq \pi^*(\theta), \forall \theta \in \Theta_1$.

Example We extend the previous example and consider $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$.

We claim that the rejection rule: Reject if $\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \geq z_\alpha$ is in fact the UMP for this test. This is because the rejection region does not depend on μ_1 and so therefore, this test, trivially, is also UMP.

Similarly, for $H_0 : \mu = \mu_0$ vs $H_a : \mu < \mu_0$ the UMP- α -level test is given by the rejection rule: Reject if $\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \leq -z_\alpha$

We first considered a simple vs simple test, and extended to a simple vs composite. We now enter the third case, composite vs composite.

Example: We may wish to test $H_0 : \mu \leq \mu_0$ vs $H_a : \mu > \mu_0$ (that is, a composite vs composite). We proceed in two steps:

1. Let $\pi(\mu)$ be the power function of the test, $T > z_\alpha$ (recall: $\pi(\mu) = P(T > z_\alpha | \mu)$, $-\infty < \mu < \infty$) and consider $H_0 : \mu \leq \mu_0$ vs $H_a : \mu > \mu_0$

We show that $\pi(\mu)$ is a non-decreasing function of μ .

$$\pi(\mu) \equiv P\left(\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_\alpha \middle| \mu\right) = P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \geq z_\alpha + \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} \middle| \mu\right) = P\left(Z \geq z_\alpha + \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} \middle| \mu\right)$$

Note that, by definition of P(Type 1 Error), $\pi(\mu_0) = \alpha$. Also note that “under μ ”, $\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ is no longer $N(0, 1)$. It is normal, but its mean is not 0.

Side Comments:

- (a) As α increases, for any $\mu > \mu_0$, $\pi(\mu, \alpha)$ increases.
- (b) If $\mu > \mu_0$, then as σ^2 increases, $\pi(\mu, \sigma^2)$ decreases.
- (c) if $\mu > \mu_0$, then as n increases, $\pi(\mu, n)$ increases.

Note that as μ increases from $-\infty$ to ∞ , $\pi(\mu)$ increases as well and if $\mu > \mu_0$, then this observation makes sense since it says that the further away μ is from μ_0 , the better the ability of the test to detect a departure from H_0

And therefore, we conclude that $\pi(\mu)$ is a non-decreasing function of μ

2. We use (1) to conclude that the rejection rule, reject if $T > z_\alpha$ is a UMP across all tests of size at most level α ($\leq \alpha$)

For this, we reason as follows:

Proposal: Use the UMP test region: Reject if $T > z_\alpha$, which was derived from $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$.

If “Nature” is thinking of some $\mu < \mu_0$, then $P(\text{Type 1 Error}) \leq \alpha$ for such μ since $\pi(\mu) \leq \pi(\mu_0)$ just shown. Therefore, our UMP α level test beats out any other test with $\mu < \mu_0$

General Approach: We use this and hope that it works (it may not)

1. Find the MP test for $H_0 : \theta = \theta_0$ vs $H_a : \theta = \theta_1$ (simple vs simple)
2. To go to $H_a : \theta > \theta_0$ or $H_a : \theta < \theta_0$, hope that R does not depend on $\mu_1 \in \Theta_1$, which therefore gives the UMP (simple vs composite)
3. To go to $H_0 : \theta \leq \theta_0$ hope to show that $\pi(\theta)$ increases for all θ to get a UMP test of size at most α (composite vs composite)

Now, consider a further alternative, $H_a : \mu \neq \mu_0$. This is called a *two-sided alternative* while $H_a : \mu > \mu_0$ is called a *one-sided alternative*.

For this two-sided alternative, consider the test statistic $T = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$. Since large values of $|T|$ should indicate a departure from H_0 in favor of H_a , it is at least reasonable (motivated from our previous discussion) to use the rejection rule: Reject H_0 in favor of H_a if $|T| \geq z_{\frac{\alpha}{2}}$.

Note: If you decide beforehand that you will do a two-sided test, then if you reject, your conclusion must be two-sided.

2.5.2 The Likelihood Ratio Test

The aim is to provide a general procedure for testing a composite $\theta \in \Theta_0$ vs a composite $\theta \in \Theta_1$. In the end, the following approach will be seen to be plausible, and moreover, it sometimes produces tests with “good” asymptotic properties.

Consider the null hypothesis, $H_0 : \theta \in \Theta_0$ vs the alternate hypothesis, $H_a : \theta \in \Theta_1$. Let $l(\theta; \underline{x})$ be the likelihood of θ at $\underline{x}$. (Recall: $l(\theta; \underline{x}) = f(\underline{x}; \theta)$ at each fixed $\underline{x}$ and θ)

Definition: The *likelihood ratio at $\underline{x}$* for testing H_0 vs H_a is defined to be:

$$\lambda(\underline{x}) = \frac{\sup_{\theta \in \Theta_0} l(\theta; \underline{x})}{\sup_{\theta \in \Theta_0 \cup \Theta_1} l(\theta; \underline{x})} = \frac{\sup_{\theta \in \Theta_0} l(\theta; \underline{x})}{\sup_{\theta \in \Theta} l(\theta; \underline{x})}$$

Note: $\lambda(\underline{x})$ is essentially the “best” explanation for the data under Θ_0 divided by the “best” explanation for the data over all θ .

Definition: The likelihood ratio test for $H_0 : \theta \in \Theta_0$ vs $H_a : \theta \in \Theta_1$ says to reject H_0 at the α -level of significance if $\lambda(\underline{x}) \leq c_\alpha$, for some c_α , determined such that:

$$\sup_{\theta \in \Theta_0} P(\lambda(\underline{x}) \leq c_\alpha) = \alpha$$

Note: Really, this probability is conditional on θ and will change for every $\theta \in \Theta_0$

It turns out that the likelihood ratio test (LRT) can be equivalently stated as:

$$\text{Reject } H_0 \text{ if } \frac{\sup_{\theta \in \Theta_0} l(\theta; \underline{x})}{\sup_{\theta \in \Theta_1} l(\theta; \underline{x})} \leq c_\alpha^*$$

But it turns out that the form with $\theta \in \Theta$ in the denominator is more convenient because we simply find $\hat{\theta}$, the MLE of θ to get $\sup_{\theta \in \Theta} l(\theta; \underline{x}) = l(\hat{\theta}; \underline{x})$, while in the latter case we would need to maximize under a constraint, which is not always easy.

The motivation for the LRT is the Neyman-Pearson lemma which says to consider the likelihood ratio in the simple vs simple case.

Example: Let $X_1, X_2, \dots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ with $\sigma^2 > 0$ unknown. Find the LRT for $H_0 : \mu = \mu_0$ vs $H_a : \mu \neq \mu_0$.

(Recall that when σ^2 is known, we saw that the test, Reject if $|\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}| > z_{\frac{\alpha}{2}}$ is not UMP, however, it is a reasonable test. Note that we now assume that σ^2 is unknown.)

Solution: First, observe that we really have a composite vs composite:

$$H_0 : (\mu, \sigma^2) \in \{(\mu, \sigma^2) : \mu = \mu_0, \sigma^2 > 0\} \text{ vs } H_a : (\mu, \sigma^2) \in \{(\mu, \sigma^2) : \mu \neq \mu_0, \sigma^2 > 0\}$$

Let $\theta = (\mu, \sigma^2)$.

So to find the LRT, we need to find:

$$\frac{\sup_{\theta \in \Theta_0} \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2} \sum (X_i - \mu)^2}}{\sup_{\theta \in \Theta} \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2} \sum (X_i - \mu)^2}}$$

First, the denominator: Here we must maximize the (Normal) likelihood over all μ and $\sigma^2 > 0$. We have seen that the MLEs are $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$ and therefore the denominator is:

$$\frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{(\hat{\sigma}^2)^{\frac{n}{2}}} e^{-\frac{1}{2\hat{\sigma}^2} \sum (X_i - \hat{\mu})^2} = \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{\hat{\sigma}^2} e^{-\frac{n}{2}}$$

Now, for the numerator: We fix $\mu = \mu_0$ and we need to find the $\hat{\sigma}^2$ that maximizes the likelihood. We have seen this to be $\hat{\sigma}^2 = \frac{1}{n} \sum (X_i - \mu_0)^2$. Therefore, the numerator is:

$$\frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{(\hat{\sigma}^2)^{\frac{n}{2}}} e^{-\frac{n}{2}}$$

And we see that the likelihood ratio is just $(\frac{\hat{\sigma}^2}{\sigma^2})^{\frac{n}{2}}$

So the LRT says to reject if $(\frac{\hat{\sigma}^2}{\sigma^2})^{\frac{n}{2}} \leq c$ for some c , or equivalently, if $\frac{\hat{\sigma}^2}{\sigma^2} \leq c^*$

Which, by taking the reciprocal of each side, we get that this is true iff: $\frac{\sigma^2}{\hat{\sigma}^2} \geq c^{**}$

Or equivalently, if and only if: $\frac{\frac{1}{n} \sum (X_i - \mu_0)^2}{\frac{1}{n} \sum (X_i - \bar{X})^2} \geq c^{**}$

That is, if $\frac{\sum (x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2}{\sum (x_i - \bar{x})^2} \geq c^{**}$

That is, if $(\frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}})^2 \geq c^{***}$

So finally, we get: $|\frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}| \geq c^{****} = c'$, and under the null hypothesis, this is just a t_{n-1} random variable.

We therefore require that the constant, c' , be such that $P(|T| > c' | \mu = \mu_0, \sigma^2) = \alpha$, so clearly, set $c = t_{n-1}(\frac{\alpha}{2})$ and finally, the LRT is: Reject H_0 if $|\frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}| > t_{n-1}(\frac{\alpha}{2})$

Comment: Unfortunately, it is often difficult to find the null distribution (under H_0) of $\lambda(\underline{x}; \theta)$. However, under quite general conditions, (smoothness, eg), the statistic $-2 \log(\lambda)$ converges to a χ^2_ν distribution. That is, the exact null distribution of $\lambda(\underline{x})$ is replaced by the approximate null distribution of $-2 \log(\lambda(\underline{x})) \sim \chi^2_\nu$

2.5.3 p-Values

Up until now, we have taken a “decision” approach to hypothesis testing. We have been forced to choose a significance level, α in advance, and a rejection rule in advance. For the same data, we cannot change either, once we have seen the data. Rather than be confined to an “arbitrary” α , it is much more common to not make a decision, but instead to provide some sort of evidence in favor of H_a (if this evidence exists). This evidence is called a *p-value*.

Let T be a given test statistic for a particular problem. In its crudest form, the p-value is the probability of observing the current value of T , T_{obs} , if the null hypothesis were true.

We reason as follows: If p is small, then there are two possible explanations:

1. p could be small because it is a very unusual observation under H_0
2. The observed value, T_{obs} , is too unusual for us to believe (1), for example, that it could have arisen under H_0 . If p is small enough (if ≤ 0.05 , we tend to say that we have strong evidence against H_0).

Note: When we say “the probability of observing the current value...” as frequentists, we mean that if we were to carry out this experiment repeatedly, the proportion of times that

we would obtain a value of T_{obs} under H_0 would be approximately p .

Details: We immediately realize a problem. If T is continuous, then $P(T = T_{obs}|\theta) = 0$ for any $\theta \in H_0$. Even if T is discrete, and n is reasonably large, $P(T = T_{obs}|\theta)$ will be very small.

Example: In tossing a fair coin 100 times, $P(T = 50)$ where T is the number of heads, is very small.

Hence the crude definition of a p -value will not work and needs modification. We need to replace T_{obs} , our reference value, by a reference set whose probability will not automatically be 0 or very small. This is what we do in practice.

The p -value is defined loosely as “The probability of observing a value as extreme or more extreme than the current T_{obs} is the null hypothesis were true.” Which leads to the following definition:

Definition: We define a p -value, $p = P(T \geq T_{obs}|\theta = \theta_0)$ for $T^* \geq T_{obs}$, $P(T \geq T^*|\theta = \theta_0) < P(T \geq T_{obs}|\theta = \theta_0)$ (i.e. if a value of T^* obtained on another day would have provided even stronger evidence against H_0) or $p = P(T \leq T_{obs}|H_0)$ otherwise.

Clearly, these definitions of the p -value are useful for one-sided tests.

We still need to define the p -value for a two-sided alternative: $H_a : \theta \neq \theta_0$. Since we cannot simultaneously observe values of T_{obs} in both tails, we define:

$$p = 2 \cdot \min \left(P(T \geq T_{obs}|\theta = \theta_0), P(T \leq T_{obs}|\theta = \theta_0) \right)$$

Essentially, we double the one-sided p -value.

In Practice:

1. Write down H_0 and H_a
2. Choose a test statistic
3. Compute your p -value
4. Simply report your p -value - let the reader give their own interpretation of the strength of the evidence against H_0

Example: Cola makers test new recipes for loss of sweetness during storage. Trained tasters rate the sweetness before and after storage. Here are the sweetness losses (sweetness before - sweetness after) found by 10 tasters for a new cola recipe.

2.0, 0.4, 0.7, 2.0, -0.4, 2.2, -1.3, 1.2, 1.1, 2.3

Is this data good evidence that the cola loses sweetness? Is this data good evidence that there is a difference in sweetness?

Note that although the problem is set up this way, you must first come up with your hypothesis and then collect the data, not the other way around.

Solution: First, it is reasonable to think of this problem as one of testing a mean, μ . Thus, let μ be the population mean difference in sweetness. We wish to test $H_0 : \mu = 0$ vs $H_a : \mu > 0$ (so that the sweetness has decreased). Next, no σ^2 is given, and so it is unknown, and further, no distribution is given, so we assume that the differences are roughly $N(\mu, \sigma^2)$.

Under this assumption, we know that the UMP test uses the test statistic $T = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{\bar{X}}{\frac{s}{\sqrt{n}}}$.

We find that $\bar{X}_{obs} = 1.02$ and $s_{obs} = 1.196$ and hence, $T_{obs} = \frac{1.02}{\frac{1.196}{\sqrt{10}}} = 2.70$.

Therefore, $p = P(T \geq 2.70 | H_0 : \mu = 0)$, but under $H_0, T \sim t_{10-1} = t_9$ and we find $P(T \geq 2.70) = 0.012$, and therefore there is strong evidence that the cola lost sweetness.

For the second question, if we were uncertain as to whether there would be a loss or gain of sweetness before seeing the data, we would conservatively have tested $H_0 : \mu = 0$ vs $H_a : \mu \neq 0$

In this case:

$$p = 2 \cdot \min \{P(T \geq 2.70 | \mu = 0), P(T \leq -2.70 | \mu = 0)\} = 2 \cdot P(T \leq -2.70) = 0.024$$

Notes:

1. The p -values give “very roughly” the proportion of times in many applications that one would obtain “such data” if the null hypothesis were true.

2. The p -value is not the probability of H_0 , nor is it the probability of H_a . It is also not either of these given the data (H_0 and H_a are not random).
3. The p -value is not the probability of a Type-1 error. This is because you can only incur an error if you make a decision. With p -values, we are not making a statistical decision.
4. The p -value is not a confidence coefficient.
5. If $p < 0.05$, do not say that “we would have rejected H_0 at the 5% level.” You have not set up a rejection region - so it does not make sense to be talking about one.
6. Many users will say, “The test was significant at the 5% level,” say, if $p < 0.05$. Therefore, in statistical context, do not use the term significant to mean substantial (that is, the every day use of the word)
7. If p is large ($p > 0.1$), then we sometimes refer to non-significance (not insignificance).
8. If p is large, then the power comes into play, even though a rejection-acceptance decision approach is not being taken. Nevertheless, the power (which does depend on a rejection region), gives some idea of how good your test would be at testing a departure from H_0 .
9. If p is small, then power is not a consideration.
10. Do not talk about testing for a “significant treatment effect.” We test for a treatment effect, or that $\mu = 3.2$ or that the median survival is greater than 5 years, etc - all about population features.
11. While you may find statistical significance in a test, this may or may not mean “clinical” significance (actual importance).
12. Beware of testing simple null hypotheses vs something else. If you have a very large sample, and your null hypothesis is too precise, you might detect a tiny difference. This is because the power increases as n increases. For example, if you are testing $H_0 : \mu = 7.3$ vs $H_a : \mu \neq 7.3$, you might find $\mu = 7.300001$, and therefore detect a departure from the null hypothesis.

3 The Bayesian Approach to Statistics

The reasoning below may be offensive to certain people. It is different from the reasoning used in frequentist statistics (as explained in the majority of these notes).

As before, we have in mind some distribution with density $f(\underline{x}; \theta)$. θ is assumed to be an unknown parameter value. The Bayesian approach is as follows:

Since θ is unknown, we can quantify our uncertainty about it through a probability distribution defined on θ .

This distribution on θ is called a *prior distribution* and is denoted by $\pi(\theta)$. Note that $\pi(\theta)$ must be specified before seeing the current data. You can use any past information (biological, physical, or past data) to construct $\pi(\theta)$. The definition of probability is not the same as that for a frequentist.

Example: Suppose that we have a coin whose probability of heads, p , say, is unknown. There are various possible $\pi(p)$'s that we could consider. If $\pi(p)$ is constant as p varies, we would say that “we have no particular preference for p .” Alternatively, if $\pi(p)$ peaks over a certain value of p (like a normal curve), then that is what we believe p to be.

Now, once we see our current data, $\underline{x}$, the Bayesian approach is to update our prior beliefs to posterior beliefs, given the data $\underline{x}$. This is done using the continuous version of Bayes' Theorem.

$$\pi(\theta|\underline{x}) = \frac{f(\underline{x}|\theta)\pi(\theta)}{m(\underline{x})}$$

Where $f(\underline{x}|\theta)$ is the likelihood of $\underline{x}$ at θ , and $m(\underline{x})$ is the marginal of $\underline{x}$, is $\int_{-\infty}^{\infty} f(\underline{x}|\theta)\pi(\theta)d\theta$

So, the aim of a Bayesian analysis is to update one's prior beliefs, $\pi(\theta)$, to posterior beliefs in the light of the currently observed data, using Bayes' Theorem.

3.1 Similarities and Differences

1. What can be done with the newly computed posterior distribution?

We could summarize the posterior distribution, $\pi(\theta|\underline{X})$, by its mean, $\mathbb{E}(\theta|\underline{x})$, or its median, or mode (value at which it attains its maximum).

2. Is there an equivalent to a confidence interval?

A $100(1 - \alpha)\%$ *posterior credibility interval* which has the property that:

$$P(L(\underline{x}) < \theta < R(\underline{x})|\underline{x}) = 1 - \alpha$$

In contrast, the frequentist confidence interval has the property that:

$$P(L(\underline{X}) < \theta < R(\underline{X})) \stackrel{\theta}{=} 1 - \alpha$$

The difference is that in the Bayesian interval, the random part is θ , and in the Frequentist interval, the random part is the $\bar{X}$.

Example: First recall that Y has a $\text{Beta}(\alpha, \beta)$ distribution if

$$f(y) = \begin{cases} \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha, \beta)} & \text{for } \alpha > 0, \beta > 0, 0 < y < 1 \\ 0 & \text{else} \end{cases}$$

where $B(\alpha, \beta) = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}dy$, the Beta function.

Note that when $\alpha = \beta = 1$, the Beta distribution reduces to the $U(0, 1)$ distribution. Also, the Beta distribution is said to be “flexible” since it can assume many different shapes, depending on the choices of α and β .

Let X be the number of observed successes in n i.i.d. Bernoulli trials with probability of success, p , given p . Suppose, that we wish, as Bayesians, to make inference about p , based on x actually observed successes.

We begin by specifying a prior distribution for p , $\pi(p)$. Suppose that we specify $p \sim \text{Beta}(\alpha, \beta)$.

We get:

$$\pi(p|x) = \frac{l(x|p)\pi(p)}{m(x)} = \frac{\binom{n}{x}p^x(1-p)^{n-x}}{m(x)} \cdot \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)} = \frac{p^{x-1+\alpha}(1-p)^{n-x-1+\beta}}{c(x; \alpha, \beta)}$$

where c is a constant that does not depend on p .

Now, we recognize that $\pi(p|x)$ is a new beta distribution, with parameters, $\alpha' = x + \alpha$ and $\beta' = n - x + \beta$ and thus $c(x; \alpha, \beta) = B_l(\alpha', \beta')$. And we see that the posterior distribution belongs to the same family as the prior distribution.

In general, we call a family of priors *conjugate* to a specific likelihood if they lead to posteriors of the same family as the prior. Conjugate families are of practical importance because of their convenience.

If $\alpha = \beta = 1$, we have a uniform prior (sometimes called a *non-informative prior*). $\pi(p|x)$ still has a Beta distribution.

3.2 Hypothesis Testing

For a Bayesian, in testing $H_0 : \theta = \theta_0$ vs $H_a : \theta = \theta_1$, we would start by imposing a prior on H_0 and H_a . That is, a prior for $P(\theta = \theta_0)$ and hence for $P(\theta = \theta_1)$. For example, maybe the prior odds are $\frac{P(H_0)}{P(H_a)} = \frac{1}{2}$.

Now, we have observed the data $\underline{x}$, and we update our prior odds to the posterior odds:

$$\frac{P(H_0|\underline{x})}{P(H_a|\underline{x})} = \frac{P(H_0|\underline{x})}{1 - P(H_0|\underline{x})} = \frac{P(\theta_0|\underline{x})}{1 - P(\theta_0|\underline{x})} = \frac{1}{100}$$