STA261 Probability and Statistics II

Haoran Yu

January 2024

Contents

1 Introduction, Sampling Distribution, and Consistency

1.1 Moment generating functions

Definition(Moments): Let X be a random variable and $c \in \mathbb{R}$ a scalar. Then: The k^{th} moment of X is:

$$\mathbb{E}\left[X^k\right]$$

and the k^{th} moment of X (about c) is:

$$\mathbb{E}\left[(X-c)^k\right]$$

Example: The first moment of X is the mean of the distribution $\mu = \mathbb{E}[X]$. This describes the center or average value.
The second moment of X about μ is the variance of the distribution $\sigma^2 = \text{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right]$. This describes the spread of a distribution (how much it varies).
The third standardized moment is called skewness $\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$ and typically tells us about the asymmetry of a distribution about its peak. If skewness is positive, then the mean is larger than the median and there are a lot of extreme high values. If skewness is negative, than the median is larger than the mean and there are a lot of extreme low values.

Definition(Moment Generating function MGF): Let X be a random variable. The moment generating function (MGF) of X is a function of a dummy variable t :

$$M_X(t) = \mathbb{E}\left[e^{tX}\right]$$

If X is discrete:

$$M_X(t) = \sum_{x \in \Omega_X} e^{tx} p_X(x)$$

If X is continuous:

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx$$

We say that the MGF of X exists, if there is a $\varepsilon > 0$ such that the MGF is finite for all $t \in (-\varepsilon, \varepsilon)$, since it is possible that the sum or integral diverges.

Example: Find the MGF of the following random variables:

(a) X is a discrete random variable with PMF:

$$p_X(k) = \begin{cases} 1/3 & k = 1 \\ 2/3 & k = 2 \end{cases}$$

Solution:

$$\begin{aligned} M_X(t) &= \mathbb{E}\left[e^{tX}\right] \\ &= \sum_x e^{tx} p_X(x) \\ &= \frac{1}{3}e^t + \frac{2}{3}e^{2t} \end{aligned}$$

(b) Y is a $\text{Unif}(0,1)$ continuous random variable.

Solution:

$$\begin{aligned} M_Y(t) &= \mathbb{E}\left[e^{tY}\right] \\ &= \int_0^1 e^{ty} f_Y(y) dy \\ &= \int_0^1 e^{ty} \cdot 1 dy, \quad [f_Y(y) = 1, 0 \le y \le 1] \\ &= \frac{e^t - 1}{t} \end{aligned}$$

Properties and uniqueness of MGFs

There are some useful properties of MGFs that we will discuss. Let X, Y be independent random variables, and $a, b \in \mathbb{R}$ be scalars. Then, recall that the moment generating function of X is: $M_X(t) = \mathbb{E}\left[e^{tX}\right]$.

(i) **Computing MGF of Linear Transformations:** We'll first see how we can compute the MGF of $aX + b$ if we know the MGF of X:

$$M_{aX+b}(t) = \mathbb{E}\left[e^{t(aX+b)}\right] = e^{tb}\mathbb{E}\left[e^{(at)X}\right] = e^{tb}M_X(at)$$

(ii) **Computing MGF of Sums:** We can also compute the MGF of the sum of independent RVs X and Y given their individual MGFs: (the third step is due to independence):

$$M_{X+Y}(t) = \mathbb{E}\left[e^{t(X+Y)}\right] = \mathbb{E}\left[e^{tX}e^{tY}\right] = \mathbb{E}\left[e^{tX}\right]\mathbb{E}\left[e^{tY}\right] = M_X(t)M_Y(t)$$

(iii) **Derivatives of MGF with respect to t:**

$$M'_X(t) = \frac{d}{dt}\mathbb{E}\left[e^{tX}\right] = \frac{d}{dt}\sum_{x\in\Omega_X} e^{tx}p_X(x) = \sum_{x\in\Omega_X}\frac{d}{dt}\left(e^{tx}p_X(x)\right) = \sum_{x\in\Omega_X} xe^{tx}p_X(x)$$

Note that if evaluate the derivative at $t = 0$, we get $\mathbb{E}[X]$ since $e^0 = 1$:

$$M'_X(0) = \sum_{x\in\Omega_X} xe^{0x}p_X(x) = \sum_{x\in\Omega_X} xp_X(x) = \mathbb{E}[X]$$

Now, let's consider the second derivative:

$$M''_X(t) = \frac{d}{dt}M'_X(t) = \frac{d}{dt}\sum_{x\in\Omega_X} xe^{tx}p_X(x) = \sum_{x\in\Omega_X}\frac{d}{dt}\left(xe^{tx}p_X(x)\right) = \sum_{x\in\Omega_X} x^2e^{tx}p_X(x)$$

If we evaluate the second derivative at $t = 0$, we get $\mathbb{E}\left[X^2\right]$:

$$M''_X(0) = \sum_{x\in\Omega_X} x^2e^{0x}p_X(x) = \sum_{x\in\Omega_X} x^2p_X(x) = \mathbb{E}\left[X^2\right]$$

(iv) **Uniqueness of MGFs:** The following are equivalent:

(a) X and Y have the same distribution.
(b) $f_X(z) = f_Y(z)$ for all $z \in \mathbb{R}$.
(c) $F_X(z) = F_Y(z)$ for all $z \in \mathbb{R}$.
(d) There is an $\varepsilon > 0$ such that $M_X(t) = M_Y(t)$ for all $t \in (-\varepsilon, \varepsilon)$ (they match on a small interval around $t = 0$).

That is M_X uniquely identifies a distribution, just like PDFs or CDFs do.

Example: Suppose $X \sim \text{Poi}(\lambda)$, meaning X has range $\Omega_X = \{0, 1, 2, \ldots\}$ and PMF:

$$p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

Compute $M_X(t)$.
Solution: First, let's recall the Taylor series:

$$e^x = \sum_{k=0}^{\infty}\frac{x^k}{k!}$$

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \sum_{k=0}^{\infty} e^{tk} p_X(k) = \sum_{k=0}^{\infty} e^{tk} \cdot e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \left(e^t\right)^k \frac{\lambda^k}{k!}$$
$$= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\left(\lambda e^t\right)^k}{k!} = e^{-\lambda} e^{\lambda e^t}$$
$$= e^{\lambda\left(e^t - 1\right)}$$

Example: If $X \sim \text{Poi}(\lambda)$, compute $\mathbb{E}[X]$ using its MGF we computed earlier $M_X(t) = e^{\lambda\left(e^t-1\right)}$.
Solution: We can prove that $\mathbb{E}[X] = \lambda$ as follows.
First we take the derivative of the moment generating function (don't forget the chain rule of calculus) and see that:

$$M_X'(t) = e^{\lambda\left(e^t-1\right)} \cdot \lambda e^t$$

Then, we know that:

$$\mathbb{E}[X] = M_X'(0) = e^{\lambda\left(e^0-1\right)} \cdot \lambda e^0 = \lambda$$

Example: If $Y \sim \text{Poi}(\gamma)$ and $Z \sim \text{Poi}(\mu)$ and Y, Z are independent, show that $Y + Z \sim \text{Poi}(\gamma + \mu)$ using the uniqueness property of MGFs.
Solution: First note that a Poi $(\gamma + \mu)$ RV has MGF $e^{(\gamma+\mu)\left(e^t-1\right)}$ (just plugging in $\gamma + \mu$ as the parameter). Since Y and Z are independent, by property introduced above,

$$M_{Y+Z}(t) = M_Y(t) M_Z(t) = e^{\gamma\left(e^t-1\right)} e^{\mu\left(e^t-1\right)} = e^{(\gamma+\mu)\left(e^t-1\right)}$$

The MGF of $Y + Z$ which we computed is the same as that of $\text{Poi}(\gamma + \mu)$. So, by the uniqueness of MGFs (which implies that an MGF can uniquely describe a distribution), $Y + Z \sim \text{Poi}(\gamma + \mu)$.

1.2 Introduction to statistic inference and random sample

Statistics inference is a branch of mathematics with wide application. It is based on probability theory and studies random phenomena based on data obtained from experiments or observations. Many problems require us to first make reasonable estimates and judgments through sampling. Statistics is the study of how to use effective methods to collect, organize and analyze data with random effects, make inferences and predictions about research problems, and provide basis and suggestions for taking certain decisions.

Statistical inference is a collection of methods that deal with drawing conclusions from data that are prone to random variation.

Definition(Population)**:** A population is a set of similar items or events which is of interest for some questions or experiment.

In statistical inference, a random subset of the population (a random sample) is chosen to represent the population in a statistical analysis.

Definition(Independent and identically distributed random variables)**:** A collection of random variables is independent and identically distributed if each random variable has the same probability distribution as the others and all are mutually independent.

Definition(Random sample of size n)**:** The collection of random variables $X_1, X_2, X_3, \ldots, X_n$ is said to be a random sample of size n if they are independent and identically distributed (i.i.d.), i.e.,

1. $X_1, X_2, X_3, \ldots, X_n$ are independent random variables, and
2. they have the same distribution, i.e,

$$F_{X_1}(x) = F_{X_2}(x) = \ldots = F_{X_n}(x), \quad \text{for all } x \in \mathbb{R}.$$

Or:

$$f_{X_1}(x) = f_{X_2}(x) = \ldots = f_{X_n}(x), \quad \text{for all } x \in \mathbb{R}.$$

Remark: For a random sample of size n, n is called the sample size.
Remark: Each random variable in the random sample from the population share the same probability distribution as the population.
Remark: The joint pdf or pmf of $X_1, \cdots, X_n$ is given by

$$f_{X_1,\ldots,X_n}(X_1, \cdots, X_n) = f(x_1) f(x_2) \cdots f(x_n) = \prod_{i=1}^{n} f(x_i)$$

Remark: Their joint CDF is given by:

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdot \ldots \cdot F_{X_n}(x_n) = \prod_{i=1}^{n} F(x_i)$$

Remark: We use capital X_i to denote the random variable in the random sample, and x_i as a specific observation from the random sample.

Definition(Statistics): Let $X_1, \cdots, X_n$ be a random sample of size n from the population and let $T(X_1, \cdots, X_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, \cdots, X_n)$. Then the random variable or random vector $Y = T(X_1, \cdots, X_n)$ is called a statistic. The only restriction for statistic is that it cannot be a function of a unknown parameters.

Example: Some examples of statistics are:

1. Sample mean:
$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i$$
2. Sample variance and sample standard deviation:
$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$
$$S = \sqrt{S^2}$$

Definition(Linear combination of random variables): Given random variables $X_1, \ldots, X_n$ and constants $a_1, \ldots, a_n$, the random variable $Y = \sum_{i=1}^{n} a_i X_i$ is a linear combination of $X_1, \ldots, X_n$

Here we have a useful lemma that is going to be used in the future proof:

Lemma: $X_1, \cdots, X_n$ be a random sample form a population, $g(x)$ be a function such that $E(g(X_1))$ and $\mathrm{Var}(g(X_1))$ exist. Then

1. $E\left(\sum_{i=1}^{n} g(X_i)\right) = n\left(E(g(X_1))\right)$
2. $\mathrm{Var}\left(\sum_{i=1}^{n} g(X_i)\right) = n\left(\mathrm{Var}(g(X_1))\right)$

Proof. We will prove this seperately:

1. Given that $X_1, X_2, \ldots, X_n$ are i.i.d., we know that $g(X_1), g(X_2), \ldots, g(X_n)$ are also i.i.d. due to the function g being applied to each X_i independently. The expectation of a sum of random variables is equal to the sum of their expectations. Thus,
$$E\left(\sum_{i=1}^{n} g(X_i)\right) = E(g(X_1)) + E(g(X_2)) + \cdots + E(g(X_n))$$
Since $X_1, X_2, \ldots, X_n$ are identically distributed, $E(g(X_1)) = E(g(X_2)) = \cdots = E(g(X_n))$. Therefore,
$$E\left(\sum_{i=1}^{n} g(X_i)\right) = nE(g(X_1))$$

2. By definition of variance, we know:

$$\begin{aligned}\operatorname{Var}\left(\sum_{i=1}^{n} g\left(X_i\right)\right) &= E\left[\sum_{i=1}^{n} g\left(X_i\right) - E\left(\sum_{i=1}^{n} g\left(X_i\right)\right]^2\right. \\ &= E\left[\sum_{i=1}^{n}\left(g\left(X_i\right) - Eg\left(X_i\right)\right)\right]^2\end{aligned}$$

Notice in the previous identity there are n terms of $\left(g\left(X_i\right) - Eg\left(X_i\right)\right)^2, i = 1, \cdots, n$, and each of them is just $\operatorname{Var}\left(g\left(X_1\right)\right)$. The remaining terms are all of the form $\left(g\left(X_i\right) - Eg\left(X_i\right)\right)\left(g\left(X_j\right) - Eg\left(X_j\right)\right), i \neq j$, which is $\operatorname{Cov}\left(g\left(X_i\right), g\left(X_j\right)\right) = 0$.

■

Lemma: Suppose that $X_1, \ldots, X_n$ are independent random variables with $E\left(X_i\right) = \mu_i$ and $\operatorname{Var}\left(X_i\right) = \sigma_i^2$. Let $Y = \sum_{i=1}^{n} a_i X_i$.

1. $E(Y) = \Sigma_{i=1}^{n} a_i \mu_i$
2. $Var(Y) = \Sigma_{i=1}^{n} a_i^2 \sigma_i^2$

Proof. We will prove this lemma:

1. The expectation of a linear combination of random variables is equal to the linear combination of their expectations. Therefore,

$$\begin{aligned}E(Y) &= E\left(\sum_{i=1}^{n} a_i X_i\right) \\ &= \sum_{i=1}^{n} E\left(a_i X_i\right)\end{aligned}$$

Since $E\left(a_i X_i\right) = a_i E\left(X_i\right)$ by the linearity of expectation, and given that $E\left(X_i\right) = \mu_i$, we have

$$= \sum_{i=1}^{n} a_i \mu_i$$

Thus, we have shown that $E(Y) = \sum_{i=1}^{n} a_i \mu_i$.

2.

$$\operatorname{Var}(Y) = \operatorname{Var}\left(\sum_{i=1}^{n} a_i X_i\right)$$

For independent random variables, this is

$$= \sum_{i=1}^{n} \operatorname{Var}\left(a_i X_i\right)$$

Applying the rule $\operatorname{Var}(aX) = a^2 \operatorname{Var}(X)$ and knowing $\operatorname{Var}\left(X_i\right) = \sigma_i^2$, we get

$$= \sum_{i=1}^{n} a_i^2 \sigma_i^2$$

Thus, we have proven that $\operatorname{Var}(Y) = \sum_{i=1}^{n} a_i^2 \sigma_i^2$.

■

Expected value and variance of statistics

Theorem: Let $X_1, \ldots, X_n$ be a random sample from a distribution (population X) with $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Then

1. $E(\bar{X}) = \mu$
2. $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$
3. $E(S^2) = \sigma^2$

Proof. We will prove this:

1. let $g(X_i) = \frac{X_i}{n}$, so $E(g(X_i)) = \frac{\mu}{n}$.
   By the first lemma, we have directly:
$$E\left(\sum_{i=1}^{n} g(X_i)\right) = n\left(E\left(g(X_1)\right)\right) = \mu$$

2. $\mathrm{Var}(g(X_i)) = \frac{\sigma^2}{n^2}$, then by Lemma:
$$\mathrm{Var}\left(\sum_{i=1}^{n} g(X_i)\right) = n\left(\mathrm{Var}\left(g(X_1)\right)\right) = n\frac{\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

3.
$$\sigma^2 = E\left[(X-\mu)^2\right]$$
   Expanding the right-hand side,
$$\sigma^2 = E\left[X^2 - 2X\mu + \mu^2\right]$$
   Applying the linearity of expectation,
$$\sigma^2 = E\left[X^2\right] - 2\mu E[X] + \mu^2$$
   Since $E[X] = \mu$, this simplifies to
$$\sigma^2 = E\left[X^2\right] - 2\mu^2 + \mu^2$$
$$\sigma^2 = E\left[X^2\right] - \mu^2$$
   Rearranging this, we get
$$E\left[X^2\right] = \sigma^2 + \mu^2$$
   So, for X_1, a random variable from the same distribution,
$$E\left(X_1^2\right) = \sigma^2 + \mu^2$$
   The sample mean $\bar{X}$ is defined as
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
   Now, let's find $E\left(\bar{X}^2\right)$:
$$E\left(\bar{X}^2\right) = E\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right)$$
$$= E\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j\right)$$

Because X_i are independent,

$$= \frac{1}{n^2}\sum_{i=1}^{n} E\left(X_i^2\right) + \frac{1}{n^2}\sum_{i\neq j} E\left(X_i\right) E\left(X_j\right)$$

Using $E\left(X_i^2\right) = \sigma^2 + \mu^2$ and $E\left(X_i\right) = \mu$,

$$\begin{aligned}
&= \frac{1}{n^2}\cdot n\left(\sigma^2+\mu^2\right) + \frac{1}{n^2}\cdot n(n-1)\mu^2 \\
&= \frac{\sigma^2}{n} + \frac{\mu^2}{n} + \frac{n(n-1)\mu^2}{n^2} \\
&= \frac{\sigma^2}{n} + \frac{n\mu^2 - \mu^2 + n(n-1)\mu^2}{n^2} \\
&= \frac{\sigma^2}{n} + \mu^2
\end{aligned}$$

So, $E\left(\bar{X}^2\right) = \frac{\sigma^2}{n} + \mu^2$.

$$\begin{aligned}
E(S^2) &= E\left(\frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]\right) \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) \quad \text{(by linearity of expectation)} \\
&= \frac{1}{n-1}\left(E\left(\sum_{i=1}^{n} X_i^2\right) - E(n\bar{X}^2)\right) \\
&= \frac{1}{n-1}\left(nE(X_1^2) - nE(\bar{X}^2)\right) \quad \text{(since } X_i \text{ are i.i.d.)} \\
&= \frac{1}{n-1}\left(n(\sigma^2+\mu^2) - nE(\bar{X}^2)\right) \quad \text{(using } E(X_1^2) = \sigma^2+\mu^2\text{)} \\
&= \frac{1}{n-1}\left(n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right) \quad \text{(since } E(\bar{X}^2) = \frac{\sigma^2}{n}+\mu^2\text{)} \\
&= \frac{1}{n-1}\left(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2\right) \\
&= \frac{1}{n-1}(n\sigma^2 - \sigma^2) \\
&= \frac{n\sigma^2 - \sigma^2}{n-1} \\
&= \frac{\sigma^2(n-1)}{n-1} \\
&= \sigma^2
\end{aligned}$$

■

1.3 Estimators and estimates

Definition(Statistical model)**:** A statistical model is a pair $(S, \mathcal{P})$, where S is the set of possible observations, i.e. the sample space, and $\mathcal{P}$ is a set of probability distributions on S.

It is assumed that there is a "true" probability distribution induced by the process that generates the observed data. We choose $\mathcal{P}$ to represent a set (of distributions) which contains a distribution that adequately approximates the true distribution.

Each possible parameter of θ determines a distribution on S; denote that distribution by F_θ. If Θ is the set of all possible parameters of θ, then $\mathcal{P} = \{F_\theta : \theta \in \Theta\}$.

Definition((Point) Estimator)**:** An estimator is a statistic, $T = g(X_1, \ldots, X_n)$, that estimates the value of a parameter of statistical model, denoted as θ.

Definition(Estimates)**:** An observed value of the statistic, $t = g(x_1, \ldots, x_n)$, is an estimate of θ.

Remark: Another notation uses a "hat" above a parameter θ to denote the estimator as $\hat{\theta}$.
Remark: Sometimes a "tilde" is used instead of a "hat" so that $\tilde{\theta}$ denotes an estimator of θ.
Example: We have some common estimators mfor statistical model:

1. Sample mean to estimate population mean.
2. Sample variance to estimate population variance.

There are infitenly many estimators for a parameter θ. So, we have some tools to determine whether we have chosen a good estimator for the parameter in question.

Definition(Biased and Unbiased)**:** Let $\hat{\theta}$ be a point estimator for a parameter θ. $\hat{\theta}$ is an unbiased estimator of θ if $E(\hat{\theta}) = \theta$. If $E(\hat{\theta}) \neq \theta, \hat{\theta}$ is said to be biased.

Definition(Bias)**:** We can also measure the bias. The bias of a point estimator $\hat{\theta}$ is given by $b(\hat{\theta}) = E(\hat{\theta}) - \theta$.

Example: By the previous lemma and proof, we know that sample mean is a unbiased estimator for population mean: $E(\bar{X}) = \mu$
Example: By the previous lemma and proof, we know that sample variance is a unbiased estimator for populatio variance: $E(S^2) = \sigma^2$

1.4 Sampling distributions

Definition(Sampling distribution)**:** The probability distribution of a statistic Y is called the sampling distribution of Y.

Definition(Standard error)**:** The standard deviation of the sampling distribution is called the standard error of that statistic.

Sampling distributions can sometimes be computed by direct computation or by approximations such as the central limit theorem.
Example: We starts by generating a hypothetical population from which samples will be drawn. This population is created to follow a normal distribution, characterized by a specified mean ($\mu = 50$) and standard deviation ($\sigma = 10$), with a total of 10,000 data points. This large population size is chosen to simulate a realistic scenario where the entire population cannot be easily measured or observed.

We then proceeds to draw 1,000 samples with replacements, each consisting of 100 observations from the population. And for each sample, we calculate the sample mean.

Then, we visualize the result with a histogram. A vertical line is added to the plot to indicate the mean of the sample means.

The Central Limit Theorem (CLT) provides the theoretical foundation for the behavior of sampling distributions. It states that, given a sufficiently large sample size, the sampling distribution of the sample mean will be approximately normally distributed, regardless of the population's distribution, with a mean equal to the population mean (μ) and a standard deviation equal to the population standard deviation (σ) divided by the square root of the sample size $(\sqrt{n})$.
Example: Suppose that the NEX automobile service center charges $\$40, \$45$, or $\$50$ for a tune-up on 4,6 , and 8 cylinder cars, with probability $0.2, 0.3$, and 0.5 , respectively. Let X be the amount of charge for a car. We choose $n = 2$ cars randomly with replacement.
Find the sampling distributions of $\bar{X}$ and S^2.
Solution: To find the sampling distributions of the sample mean $\bar{X}$ and the sample variance S^2 for $n = 2$ cars chosen randomly with replacement from the given distribution of charges and probabilities, we first define the possible charges and their probabilities:

- Charges: $\$40, \$45, \$50$
- Probabilities: $0.2, 0.3, 0.5$, respectively.

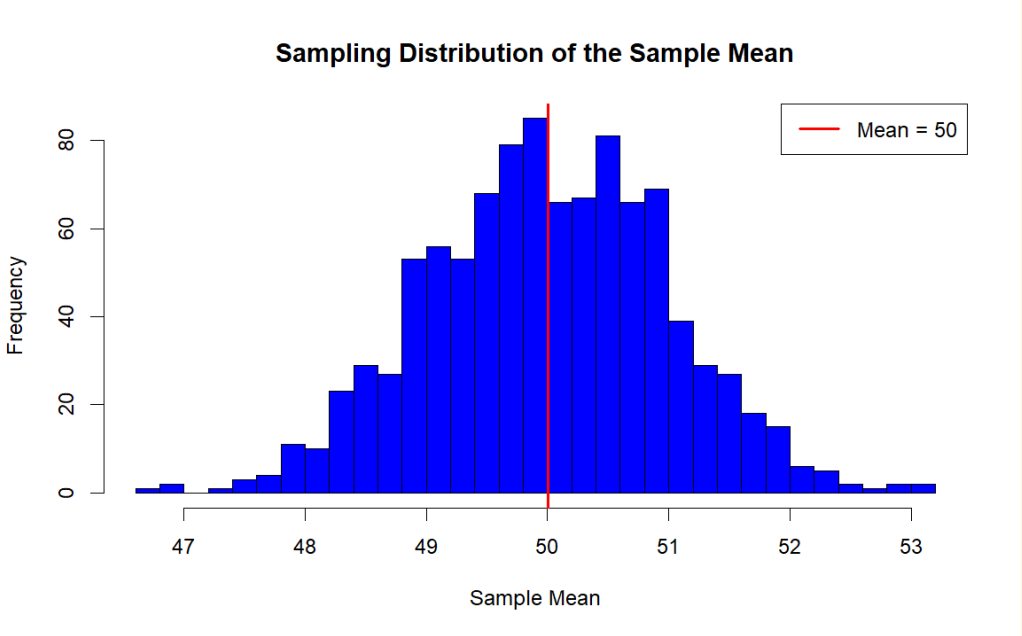


Figure 1: Sample distribution example

Let X represent the charge for a tune-up, with X taking values in $\{40, 45, 50\}$ with probabilities $P(X = 40) = 0.2, P(X = 45) = 0.3$, and $P(X = 50) = 0.5$.
Since we are drawing two cars with replacement, we calculate $\bar{X}$ for each possible pair of samples. The possible pairs (with replacement) are $(40, 40), (40, 45), (40, 50), (45, 40), (45, 45), (45, 50), (50, 40), (50, 45)$, and $(50, 50)$.
The sample variance S^2 for each pair is calculated using the formula:

$$S^2 = \frac{\sum \left(X_i - \bar{X}\right)^2}{n - 1}$$

where n is the sample size (2 in this case), X_i are the sample values, and $\bar{X}$ is the sample mean.
For each pair, we'll calculate $\bar{X}$ and S^2, and then determine the probability of each $\bar{X}$ and S^2 occurring by multiplying the probabilities of drawing each car type together.
And we have:

$$\bar{X} = \$40.00 \text{ with probability } 0.04$$
$$\bar{X} = \$42.50 \text{ with probability } 0.12$$
$$\bar{X} = \$45.00 \text{ with probability } 0.29$$
$$\bar{X} = \$47.50 \text{ with probability } 0.30$$
$$\bar{X} = \$50.00 \text{ with probability } 0.25$$

$$S^2 = \$0.00 \text{(no variance within the sample) with probability } 0.38$$
$$S^2 = \$12.50 \text{with probability } 0.42$$
$$S^2 = \$50.00 \text{with probability } 0.20$$

Then the probability distributions of $\bar{X}$ and S^2 become

$\bar{x}$	40	42.5	45	47.5	50	s^2	0	12.5	50
$p_{\bar{X}}(\bar{x})$	0.04	0.12	0.29	0.30	0.25	$p_S^2(s^2)$	0.38	0.42	0.20

1.5 Consistency

We will denote estimators as $\hat{\theta}_n$ to be explicit about their dependence on the sample size n.

The point estimator is a function of the sample, so the point estimator is still a random variable. Under the conditions of a certain sample size, we cannot require it to be completely equivalent to the true value of the unknown parameter, but if the sample size continues to increase, it can get closer and closer to the true value of the unknown parameter, the intensity (probability) of controlling the value near the true value becomes larger and larger, then this is a good estimator. This property is called consistency.

Definition(Consistent estimator): $\hat{\theta}_n$ is a consistent estimator of θ if, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\hat{\theta}_n - \theta\right| \leq \varepsilon\right) = 1$$

or, equivalently,

$$\lim_{n\to\infty} P\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) = 0$$

Consistency is often denoted as $\hat{\theta}_n \to^p \theta$.

Sufficient condition for conssitent unbiased estimator

Theorem: Not all unbioased estimator is consistent. An unbiased estimator $\hat{\theta}_n$ for θ is consistent for θ if

$$\lim_{n\to\infty} \operatorname{Var}\left(\hat{\theta}_n\right) = 0$$

Proof. By definition, $\hat{\theta}_n$ is unbiased, so $E\left[\hat{\theta}_n\right] = \theta$.
We are given that $\lim_{n\to\infty} \operatorname{Var}\left(\hat{\theta}_n\right) = 0$.
With Chebyshev's inequality and the variance condition, for any $\epsilon > 0$,

$$P\left(\left|\hat{\theta}_n - \theta\right| \geq \epsilon\right) \leq \frac{\operatorname{Var}\left(\hat{\theta}_n\right)}{\epsilon^2}$$

As $n \to \infty, \operatorname{Var}\left(\hat{\theta}_n\right) \to 0$, making the right-hand side of the inequality approach 0.
If the probability of the difference between $\hat{\theta}_n$ and θ being greater than or equal to ϵ approaches 0 , then the probability of the difference being less than ϵ approaches 1 , which implies

$$\lim_{n\to\infty} P\left(\left|\hat{\theta}_n - \theta\right| < \epsilon\right) = 1$$

This shows that $\hat{\theta}_n$ converges in probability to θ, making it a consistent estimator for θ by definition.
Therefore, we have proved that an unbiased estimator $\hat{\theta}_n$ for θ is consistent for θ if

$$\lim_{n\to\infty} \operatorname{Var}\left(\hat{\theta}_n\right) = 0.$$

■

Example: Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and finite variance σ^2. Show that $\bar{X}_n$ is a consistent estimator of μ.
Solution: The sample mean $\bar{X}_n$ is defined as:

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

The expectation of $\bar{X}_n$ is:

$$E\left[\bar{X}_n\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right]$$

Since each X_i has the same expectation μ, this simplifies to:

$$E\left[\bar{X}_n\right] = \frac{1}{n} \cdot n\mu = \mu$$

This shows that $\bar{X}_n$ is an unbiased estimator of μ. We did a slightly different proof as the theorem above in the previous section. The variance of $\bar{X}_n$ is:

$$\operatorname{Var}\left(\bar{X}_n\right) = \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{Var}\left(X_i\right)$$

Assuming that the X_i are independent and identically distributed (i.i.d.) with variance σ^2, this becomes:

$$\operatorname{Var}\left(\bar{X}_n\right) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$$

We did a similar proof on the same thing in the previous sectionas well.
For $\bar{X}_n$ to be a consistent estimator of $\mu, \bar{X}_n$ must converge in probability to μ as $n \to \infty$. This means that for any $\epsilon > 0$, the probability that the absolute difference between $\bar{X}_n$ and μ is less than ϵ approaches 1 as n increases:

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

Using Chebyshev's inequality, as previously explained, and noting that $E\left[\bar{X}_n\right] = \mu$ and $\operatorname{Var}\left(\bar{X}_n\right) = \sigma^2/n$, we have:

$$P\left(\left|\bar{X}_n - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}$$

As $n \to \infty, \frac{\sigma^2}{n\epsilon^2} \to 0$, which implies that $P\left(\left|\bar{X}_n - \mu\right| < \epsilon\right) \to 1$, satisfying the condition for consistency.
Therefore, we have shown that $\bar{X}_n$ is a consistent estimator of μ.

Weak law of large numbers

The example above that says $\bar{X}_n$ is consistent for μ, or converges in probability to μ is called Weak law of large numbers (WLLN).

Theorem: Given a collection of iid samples from a random variable with finite mean, the sample mean converges in probability to the expected value

$$\bar{X}_n \xrightarrow{P} \mu \quad \text{when } n \to \infty.$$

That is, for any positive number ε,

$$\lim_{n\to\infty} \Pr\left(\left|\bar{X}_n - \mu\right| < \varepsilon\right) = 1.$$

Weak law of large numbers can be visualized with Bernoulli experiment. Consider flipping a sequence of identical fair coins. Let $\bar{X}_n$ be the fraction of the first n coins that are heads. Then $\bar{X}_n = (X_1 + \cdots + X_n)/n$, where $X_i = 1$ if the i th coin is heads, otherwise $X_i = 0$.

Using R, we will generate a sequence of coin flips for a large number of trials, where each coin flip is represented by a random variable X_i that takes the value 1 for heads and 0 for tails. And we plot the value of $\bar{X}_n$ as n increases to show how it converges to the expected value of 0.5 (since the coin is fair, the probability of heads, $p = 0.5$).

Preserving consistency as functions

Theorem: Suppose that $\hat{\theta}_n$ is a consistent estimator of θ and that $\hat{\theta}'_n$ is a consistent estimator of θ'. Then

(1) $\hat{\theta}_n \pm \hat{\theta}'_n$ is a consistent estimator of $\theta \pm \theta'$.

(2) $\hat{\theta}_n \times \hat{\theta}'_n$ is a consistent estimator of $\theta \times \theta'$.

(3) If $\theta' \neq 0, \frac{\hat{\theta}_n}{\hat{\theta}'_n}$ is a consistent estimator of $\frac{\theta}{\theta'}$.

(4) If $g(\cdot)$ is a real-valued function that is continuous at θ, then $g\left(\hat{\theta}_n\right)$ is a consistent estimator of $g(\theta)$. This is also known as continuous mapping theorem

Proof. We will prove the parts individually:

(1) Given $\hat{\theta}_n$ is a consistent estimator of θ and $\hat{\theta}'_n$ is a consistent estimator of θ'.

By the definition of consistency, $\hat{\theta}_n \xrightarrow{p} \theta$ and $\hat{\theta}'_n \xrightarrow{p} \theta'$.

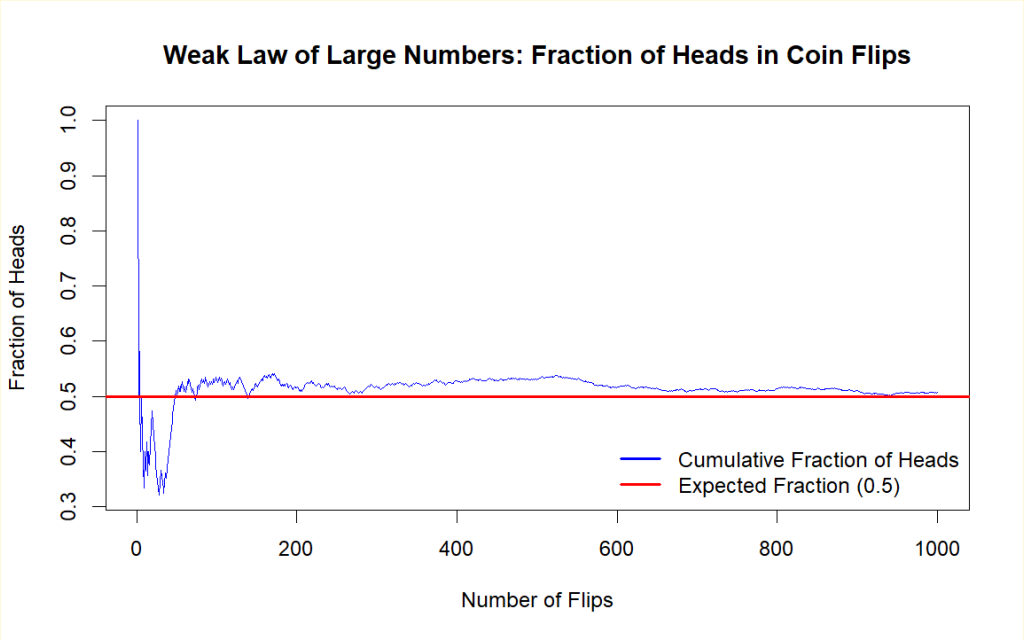


Figure 2: Visualize WLLN

The sum (or difference) of two convergent sequences of random variables also converges to the sum (or difference) of their limits. Thus, $\hat{\theta}_n \pm \hat{\theta}'_n \xrightarrow{p} \theta \pm \theta'$.

(2) The product of two sequences of random variables that converge in probability to some limits will converge in probability to the product of those limits. Thus, if $\hat{\theta}_n \xrightarrow{p} \theta$ and $\hat{\theta}'_n \xrightarrow{p} \theta'$, then $\hat{\theta}_n \times \hat{\theta}'_n \xrightarrow{p} \theta \times \theta'$.

(3) Assuming $\theta' \neq 0$ and $\hat{\theta}'_n \xrightarrow{p} \theta'$, the probability that $\hat{\theta}'_n$ equals zero goes to zero as n approaches infinity.

The Slutsky's Theorem allows for the division, stating that if $\hat{\theta}_n \xrightarrow{p} \theta$ and $\hat{\theta}'_n \xrightarrow{p} \theta'$, then $\frac{\hat{\theta}_n}{\hat{\theta}'_n} \xrightarrow{p} \frac{\theta}{\theta'}$, assuming $\theta' \neq 0$.

(4) $\hat{\theta}_n$ is consistent for θ implies that for every $\epsilon > 0, \lim_{n\to\infty} P\left(\mid \hat{\theta}_n - \theta \mid < \epsilon\right) = 1$.

g being continuous at θ means that for every $\epsilon' > 0$, there exists a $\delta > 0$ such that if $|x - \theta| < \delta$, then $|g(x) - g(\theta)| < \epsilon'$.

By the definition of consistency, for any $\delta > 0$, there exists an N such that for all $n \geq N, P\left(\left|\hat{\theta}_n - \theta\right| < \delta\right) > 1 - \epsilon$, where ϵ is an arbitrary small positive number representing our tolerance level for the probability of deviation.

Given the continuity of g at θ, choose δ corresponding to a given ϵ' such that if $\mid \hat{\theta}_n - \theta \mid < \delta$, then $\left|g\left(\hat{\theta}_n\right) - g(\theta)\right| < \epsilon'$.

Since $\hat{\theta}_n \xrightarrow{p} \theta$, for this chosen δ, as n approaches infinity, the probability that $\mid \hat{\theta}_n - \theta \mid < \delta$ approaches 1 . Therefore, the probability that $\left|g\left(\hat{\theta}_n\right) - g(\theta)\right| < \epsilon'$ also approaches 1 , by the continuity of g.

This implies that for any $\epsilon' > 0$,

$$\lim_{n\to\infty} P\left(\left|g\left(\hat{\theta}_n\right) - g(\theta)\right| < \epsilon'\right) = 1,$$

demonstrating that $g\left(\hat{\theta}_n\right)$ converges in probability to $g(\theta)$. Hence, $g\left(\hat{\theta}_n\right)$ is a consistent estimator for $g(\theta)$, as required.

■

2 Distribution theory

2.0.1 Normal Distribution theory

Distribution of linear combination of normal RVs

Theorem: Let $X_1, \ldots, X_n$ be independent random variables from a Normal distribution with $E(X_i) = \mu_i$ and $\mathrm{Var}(X_i) = \sigma_i^2$. Then

$$Y = \sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

Proof. The characteristic function of a random variable X is defined as $\phi_X(t) = E\left[e^{itX}\right]$, where i is the imaginary unit and $E[\cdot]$ denotes the expected value. For a normal random variable $X \sim N(\mu, \sigma^2)$, the characteristic function is given by

$$\phi_X(t) = e^{it\mu - \frac{1}{2}\sigma^2 t^2}.$$

For the linear combination $Y = \sum_{i=1}^{n} a_i X_i$, we want to find its characteristic function. Since X_i are independent, the characteristic function of Y is the product of the characteristic functions of $a_i X_i$. For each term $a_i X_i$, its characteristic function is derived from the characteristic function of X_i by replacing t with $a_i t$ and adjusting for the mean and variance of X_i :

$$\phi_{a_i X_i}(t) = e^{ita_i\mu_i - \frac{1}{2}(a_i^2\sigma_i^2)t^2}.$$

The characteristic function of Y is therefore

$$\phi_Y(t) = \prod_{i=1}^{n} \phi_{a_i X_i}(t) = \prod_{i=1}^{n} e^{ita_i\mu_i - \frac{1}{2}(a_i^2\sigma_i^2)t^2} = e^{\sum_{i=1}^{n}\left(ita_i\mu_i - \frac{1}{2}(a_i^2\sigma_i^2)t^2\right)}.$$

Simplifying the exponent, we get

$$\phi_Y(t) = e^{it\sum_{i=1}^{n} a_i\mu_i - \frac{1}{2}t^2\sum_{i=1}^{n} a_i^2\sigma_i^2}.$$

Since the characteristic function uniquely determines the distribution, we conclude that

$$Y = \sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right).$$

■

Example: If $X_1, \ldots, X_n$ are random sample (iid) from $N(\mu, \sigma^2)$, then

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Proof. For the sample mean, we have a specific linear combination where each $a_i = \frac{1}{n}, \mu_i = \mu$, and $\sigma_i^2 = \sigma^2$ for all i. Therefore, the sample mean can be represented as:

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

Applying the theorem, the mean of this linear combination is:

$$E[\bar{X}] = \sum_{i=1}^{n} a_i\mu_i = \sum_{i=1}^{n} \frac{1}{n}\mu = \mu.$$

The variance of this linear combination is:

$$\mathrm{Var}(\bar{X}) = \sum_{i=1}^{n} a_i^2\sigma_i^2 = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 \sigma^2 = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$$

By applying the theorem, we have shown that $\bar{X}$, the sample mean of n iid normal random variables $X_1, \ldots, X_n$ with mean μ and variance σ^2, is also normally distributed with mean μ and variance $\frac{\sigma^2}{n}$. This directly leads us to conclude:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$

■

Example: Amount dispensed (in ounces) by a beer bottling machine is normally distributed with $\sigma^2 = 1$. For a sample of size $n = 9$,

1. Find the probability that the sample mean is within 0.3 onces of μ.

   Solution: The sampling distribution of the sample mean $\bar{X}$ for a normally distributed population is also normally distributed with mean μ and standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

   Therefore, $\sigma_{\bar{X}} = \frac{1}{\sqrt{9}} = \frac{1}{3}$.

   To find the probability that the sample mean is within 0.3 ounces of μ, we standardize this interval to use the standard normal distribution Z, where

   $$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}.$$

   The probability we are looking for can be expressed in terms of Z as:

   $$\begin{aligned} &P(\mu - 0.3 < \bar{X} < \mu + 0.3) \\ &= P\left(-\frac{0.3}{\sigma_{\bar{X}}} < Z < \frac{0.3}{\sigma_{\bar{X}}}\right) \\ &= P(-0.9 \leq Z \leq 0.9) = 1 - 2P(Z > 0.9) = 1 - 2(0.184) = 0.632 \end{aligned}$$

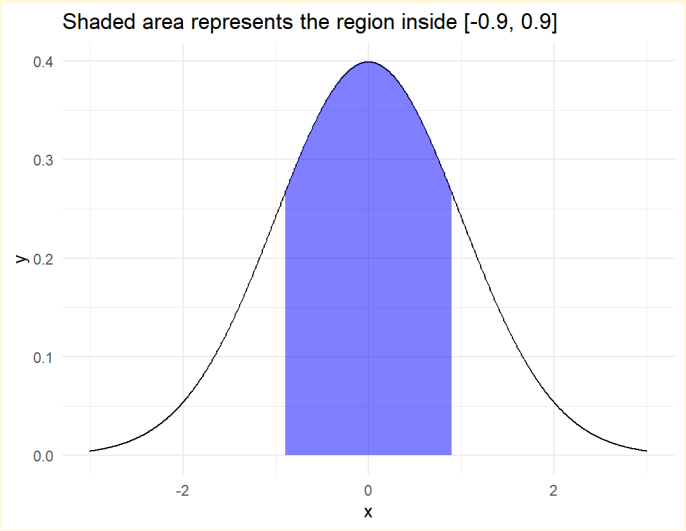


Figure 3: Area between -0.9 and 0.9

2. How big of a sample size do we need if we want the sample mean to be within 0.3 ounces of μ with probability 0.95?

   Solution: we need to find n such that:

   $$P(-0.3\sqrt{n} \leq Z \leq 0.3\sqrt{n}) = 0.95$$

   From the standard normal table, we see that

   $$P(-1.96 \leq Z \leq 1.96) = 0.95$$

   Hence $0.3\sqrt{n} = 1.96$ or $n = 42.68 \approx 43$.

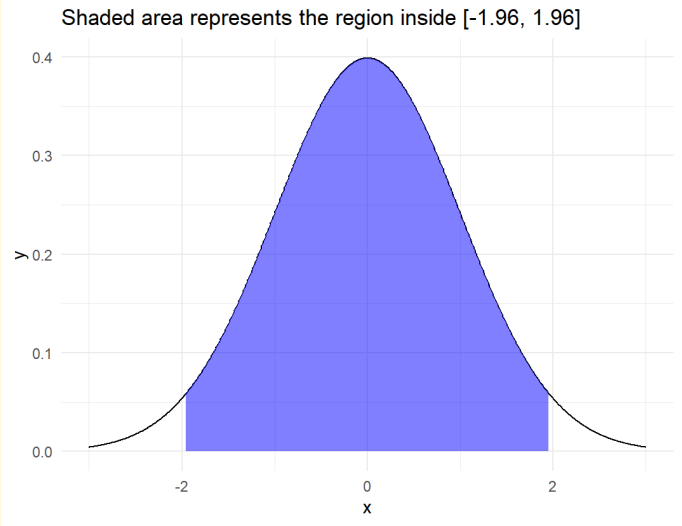


Figure 4: Area between -1.96 and 1.96

The central limit theorem

Theorem: Let $X_1, \ldots X_n$ be a sequence of independent and identically distributed random variables with mean μ and (finite) variance σ^2. Then, the standardized sample mean approaches the standard Normal distribution:

$$\text{As } n \to \infty, \quad \bar{Z}_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \to N(0,1)$$

To illustrate the central limit theorem, we draw $1, 2, 3, 5, 10, 30$ samples from a standard Uniform distribution. Repeat this 1000 times, each time compute the sample mean. And then Create the histograms in R:

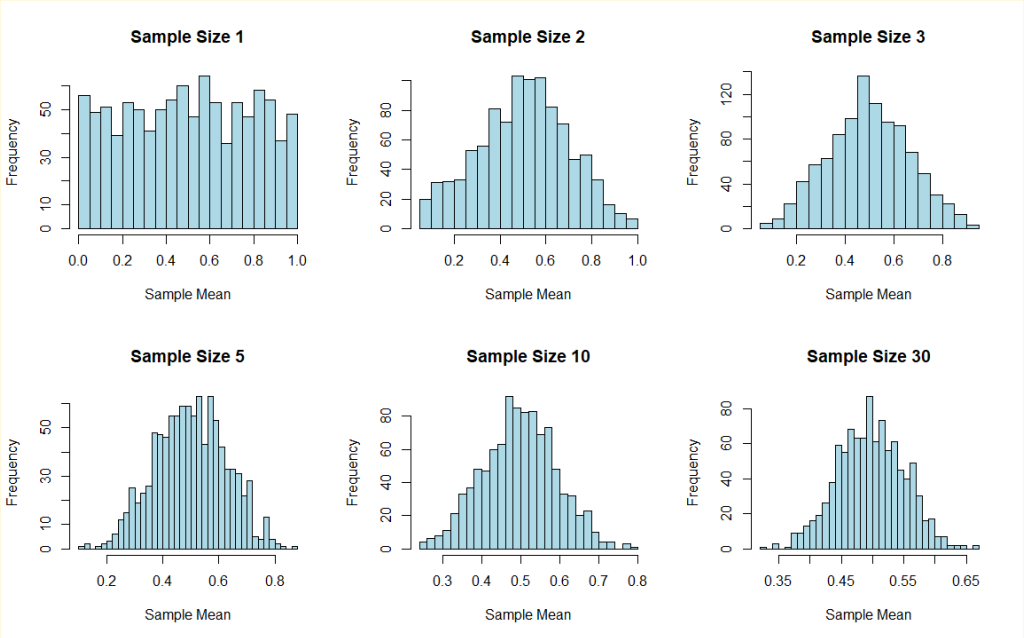


Figure 5: CLT simulation

These histograms visually demonstrate the CLT by showing that even though the original distribution (Uniform in this case) is not normal, the distribution of the sample means tends towards a normal distribution as the sample size increases.

The theorem that $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ applies specifically to samples from a normally distributed population. It states that the sample mean itself follows a normal distribution, with its mean equal to the population mean and its variance equal to the population variance divided by the sample size, regardless of the sample size n.

The CLT, on the other hand, is more general and does not require the original population to be normally distributed. It states that as the sample size n becomes large, the distribution of the standardized sample mean $\bar{Z}_n$ approaches a standard normal distribution ($\mathcal{N}(0,1)$), regardless of the population's original distribution, provided the population has a finite mean and variance.

Example: The service times for customers coming through a Navy Exchange checkout counter are iid with $\mu = 1.5$ and $\sigma = 1.0$. Approximate the probability that $n = 100$ customers can be served in less than 2 hours.

Solution: Let X_i be the service time for the i th customer, then we have

$$P\left(\sum_{i=1}^{100} X_i \leq 120\right) = P(\bar{X} \leq 1.2)$$

Since the sample size is large enough, using the CLT, $\bar{X}$ is approximately normally distributed with mean $\mu_{\bar{X}} = \mu = 1.5$ and $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} = 1/100$. Hence

$$P(\bar{X} \leq 1.2) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{1.2 - 1.5}{1/\sqrt{100}}\right) \approx P(Z \leq -3) = 0.0013$$

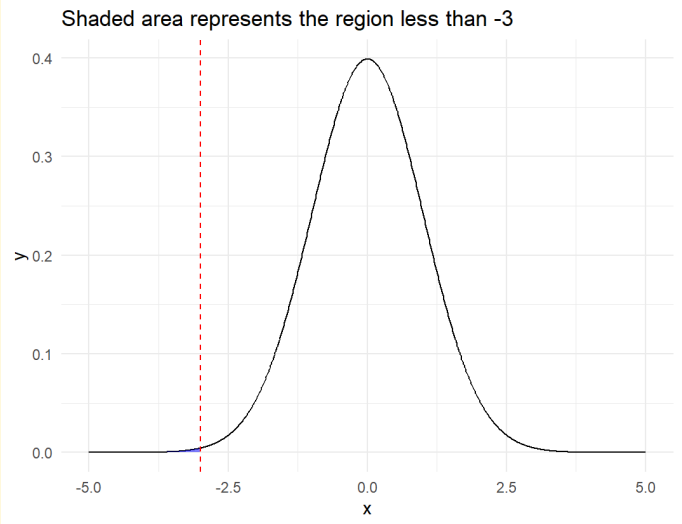


Figure 6: Area less than -3

Convergence in normal distribution

Theorem: Suppose that U_n has a distribution function that converges to a standard normal distribution as $n \to \infty$. If W_n converges in probability to 1 , then $\frac{U_n}{W_n}$ converges in distribution to a standard normal distribution.

Example: Suppose $X_1, X_2, \ldots, X_n$ are random sample from a distribution with mean μ and variance σ^2. Show that the distribution function of $Z_n = \frac{\bar{X}_n - \mu}{S/\sqrt{n}}$ converges to a standard normal distribution.

Solution: First, from the previous example, S_n^2 is consistent estimator of σ^2 which implies that $S = \sqrt{S_n^2}$ is a consistent estimator of σ (even though, it is a biased estimator). That is, $\frac{S}{\sigma}$ conveges to 1 in probability. Secondly, by the CLT, $U_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ converges in distribution to a standard normal distribution. Thus, by the previous Thorem, as $n \to \infty$,

$$Z_n = \frac{\sqrt{n}\left(\bar{X}_n - \mu\right)/\sigma}{S/\sigma} = \frac{\bar{X}_n - \mu}{S/\sqrt{n}}$$

converges in distribution to a standard normal distribution.

2.0.2 Chi-Square distribution

Definition(Chi-Square distribution): If $Z_1, \ldots, Z_k$ are independent, standard normal random variables, then the sum of their squares,

$$Q = \sum_{i=1}^{k} Z_i^2,$$

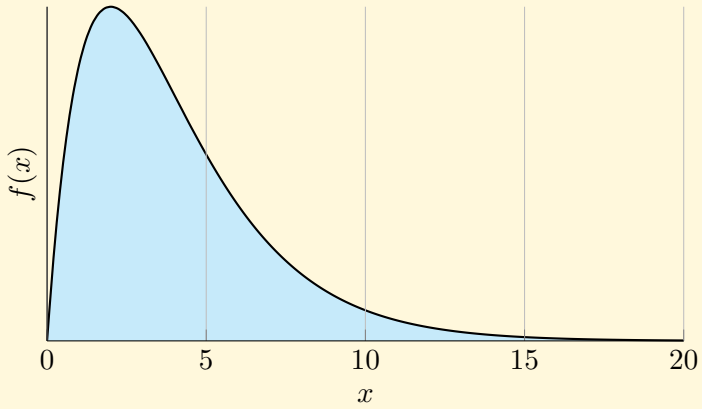
is distributed according to the chi-squared distribution with k degrees of freedom. This is usually denoted as

$$Q \sim \chi^2(k) \text{ or } Q \sim \chi_k^2.$$

The probability density function (pdf) of the chi-squared distribution is

$$f(x)_k = \begin{cases} \frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma\left(\frac{k}{2}\right)}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

where $\Gamma(k/2)$ denotes the gamma function, which has closed-form values for integer k.
Here we have a example Chi-square distribution with $k = 4$:



Distributions of sum of squared of RVs from standard normal

Theorem: Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$. Then $Z_i = \frac{X_i - \mu}{\sigma}$ are iid random variables from the standard normal distribution and

$$Z_i^2 \sim \chi_{(1)}^2, \quad i = 1, \ldots, n.$$

Further,

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi_{(n)}^2$$

Proof. Given $X_i \sim N(\mu, \sigma^2)$, we define the transformation $Z_i = \frac{X_i - \mu}{\sigma}$.
The transformation Z_i standardizes X_i, resulting in:

(i) $E[Z_i] = E\left[\frac{X_i - \mu}{\sigma}\right] = \frac{1}{\sigma}(E[X_i] - \mu) = 0$

(ii) $\text{Var}(Z_i) = \text{Var}\left(\frac{X_i - \mu}{\sigma}\right) = \frac{1}{\sigma^2}\text{Var}(X_i) = 1.$

Thus, Z_i is a standard normal variable, $Z_i \sim N(0, 1)$.
Squaring a standard normal variable, Z_i^2, yields a variable that follows a Chi-square distribution with 1 degree of freedom $\left(X_{(1)}^2\right)$. This is by definition of the Chi-square distribution, which is the sum of the squares of k independent standard normal variables. Here, each Z_i^2 is essentially the case of $k = 1$.
The sum of squared standard normal variables, $\sum_{i=1}^{n} Z_i^2$, is the sum of n independent Chi-square variables each with 1 degree of freedom. By the properties of the Chisquare distribution, this sum:

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2$$

follows a Chi-square distribution with n degrees of freedom, denoted as $X_{(n)}^2$. ■

Definition(Quatiles of Chi-Square distribution)**:** For a χ^2 distribution with k degrees of freedom, the p-th quantile x_p is defined by:

$$P\left(X \leq x_p\right) = p$$

where X is a χ^2-distributed random variable with k degrees of freedom, and p is a probability such that $0 < p < 1$. For example, for freedom 10:

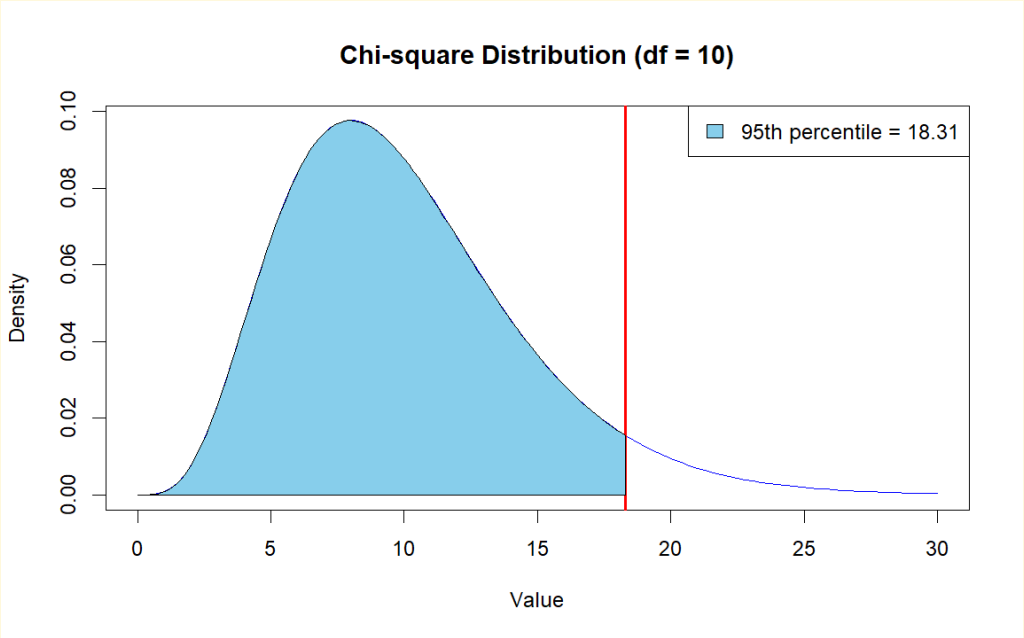


Figure 7: example 95th percentile

Example: Let $Z_1, Z_2, \ldots, Z_6$ be a random sample from the standard Normal distribution. Find the number b such that $P\left(\sum_{i=1}^{6} Z_i^2 \leq b\right) = 0.95$.

Solution: Since $\sum_{i=1}^{6} Z_i^2 \sim X^2_{(6)}$, from the Chi-squared Table, we have

$$P\left(\sum_{i=1}^{6} Z_i^2 > 12.59\right) = 0.05 \text{ or } P\left(\sum_{i=1}^{6} Z_i^2 \leq 12.59\right) = 0.95$$

Hence $b = 12.5916$ is the 0.95 quantile (95th percentile) of the sum of the squares of 6 independent random variables from the standard normal distribution.

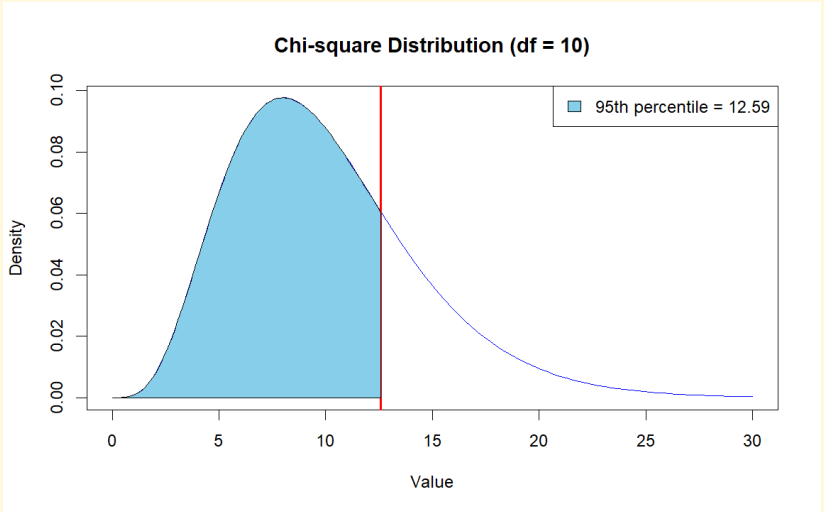


Figure 8: 95th percentile

The marginal distribution of S^2

Theorem: Let $X_1, X_2, \ldots, X_n$ be a random sample from a Normal distribution with mean μ and variance σ^2. Then

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

Furthermore, $\bar{X}$ and S^2 are independent random variables.

Example: In the example, suppose that the amount dispensed (in ounces) has a normally distribution with $\sigma^2 = 1$. For a sample of size $n = 10$, find b_1 and b_2 such that $P\left(b_1 \leq S^2 \leq b_2\right) = 0.9$

Solution: From the theorem, we see that

$$P\left(b_1 \leq S^2 \leq b_2\right) = P\left(\frac{(n-1)b_1}{\sigma^2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \frac{(n-1)b_2}{\sigma^2}\right)$$

Since $(n-1)S^2/\sigma^2 \sim X^2_{(n-1)}$, where $n = 10$ and $\sigma^2 = 1$. From the Chi-squared table, we need to find a_1 and a_2 such that $P\left(a_1 \leq (n-1)S^2 \leq a_2\right) = 0.9$.
One solution for a_1 and a_2 are 5th and 95th percentiles of $X^2_{(9)}$ are 3.325 and 16.919, respectively. Hence $b_1 = 0.369$ and $b_2 = 1.880$.

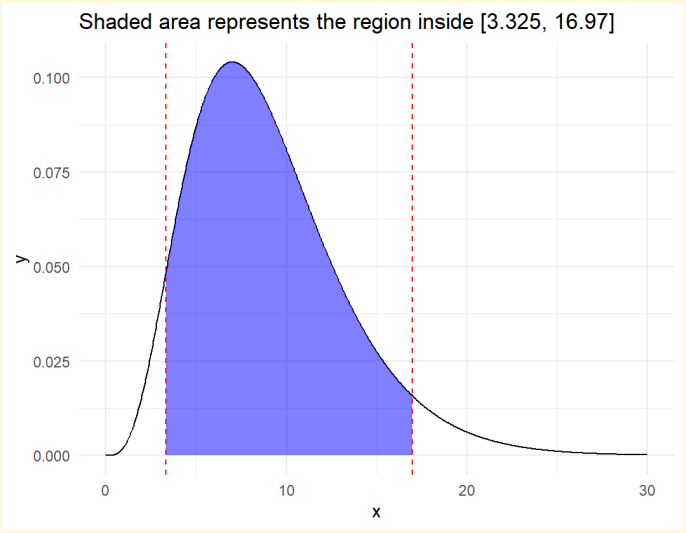


Figure 9: Area between 5th and 95th percentile

2.1 t distribution

Definition(t-distribution): Random variable T has a t distribution with p degrees of freedom, written as $T \sim t_p$ if it has pdf

$$f_T(t) = \frac{\Gamma\left(\frac{p-1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)} \frac{1}{(p\pi)^{1/2}} \frac{1}{(1 + t^2/p)^{(p+1)/2}}$$

Definition(Variable with t-distribution): Let Z be random variable from $N(0,1)$ and U be a random variable from $X^2_{(n)}$. If Z and U are independent, then

$$T = \frac{Z}{\sqrt{U/n}} \sim t_n$$

where t_n represents the $\mathbf{t}$ distribution with n degrees of freedom.

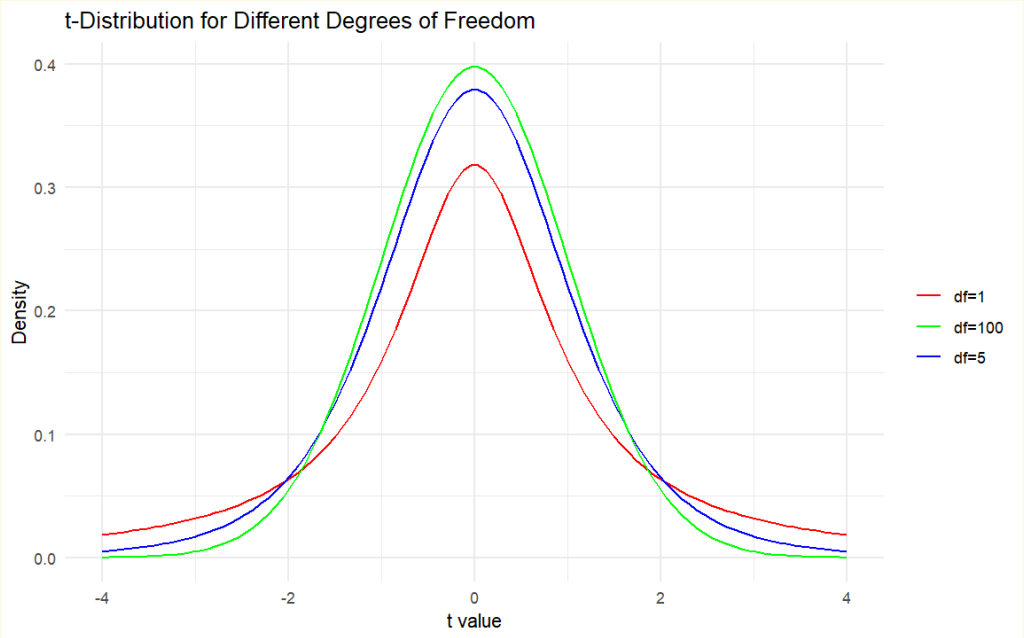


Figure 10: t-distribution example

When the n gets sufficient large, it approaches the standard normal distribution:

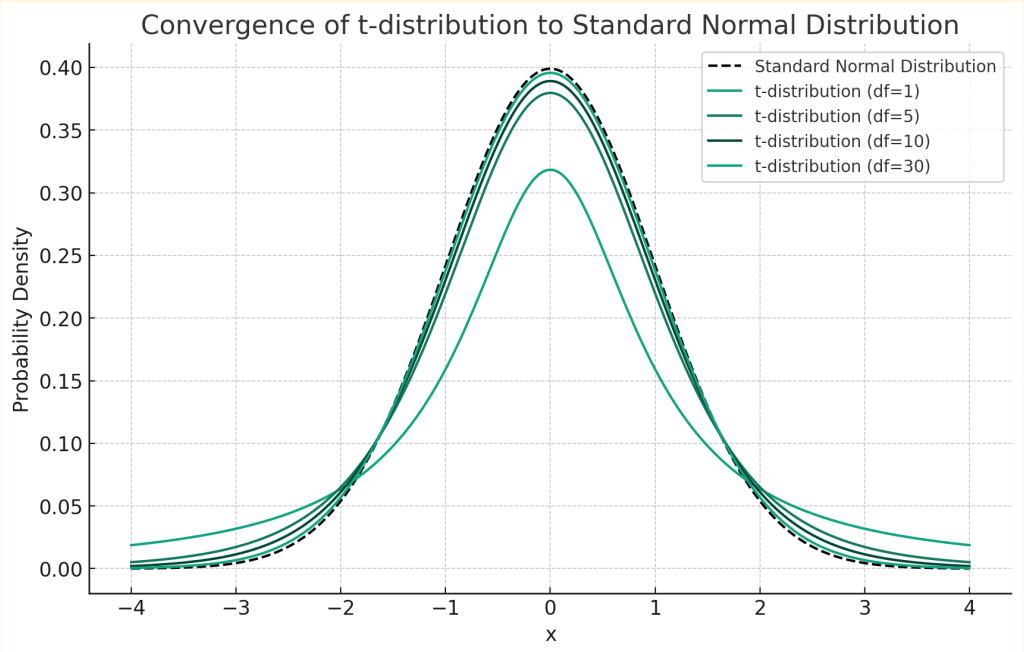


Figure 11: Demostrating t appraoches standard normal

Example: Let $X_1, \ldots, X_n$ are iid (random sample) from $N(\mu, \sigma^2)$, then

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{\left[(n-1)S^2/\sigma^2\right]/(n-1)}} = \frac{Z}{\sqrt{U/(n-1)}} \sim t_{n-1}$$

The α quantile of the t_n distribution is denoted as $t_\alpha(n)$, which means: for a given $\alpha(0 < \alpha < 1)$, when $T \sim t_n$, there is

$$P\left(T \leqslant t_\alpha(n)\right) = \alpha.$$

The value $t_\alpha(n)$ can be acquired from the t-distribution table. Because t_n is symmetric, it means:

$$t_\alpha(n) = -t_{1-\alpha}(n)$$

Example: The tensile strength of a type of wire is normally distributed with unknown mean μ and variance σ^2. Six pieces are randomly selected from a large roll and the tensile strength are measured. Find the probability that $\bar{X}$ will be within $2S/\sqrt{n}$ of the true population mean μ.

Solution: We need to find

$$P\left(-\frac{2S}{\sqrt{n}} \leq \bar{X} - \mu \leq \frac{2S}{\sqrt{n}}\right) = P\left(-2 \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq 2\right) = P(-2 \leq T \leq 2)$$

Because $n = 6$, from the previous example, we have $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_5$ and from the t table, we obtain

$$P(-2.015 \leq T \leq 2.015) = 0.90$$

Hence, the probability that $\bar{X}$ will be within 2 estimated standard deviation of μ is slightly less than 0.90.

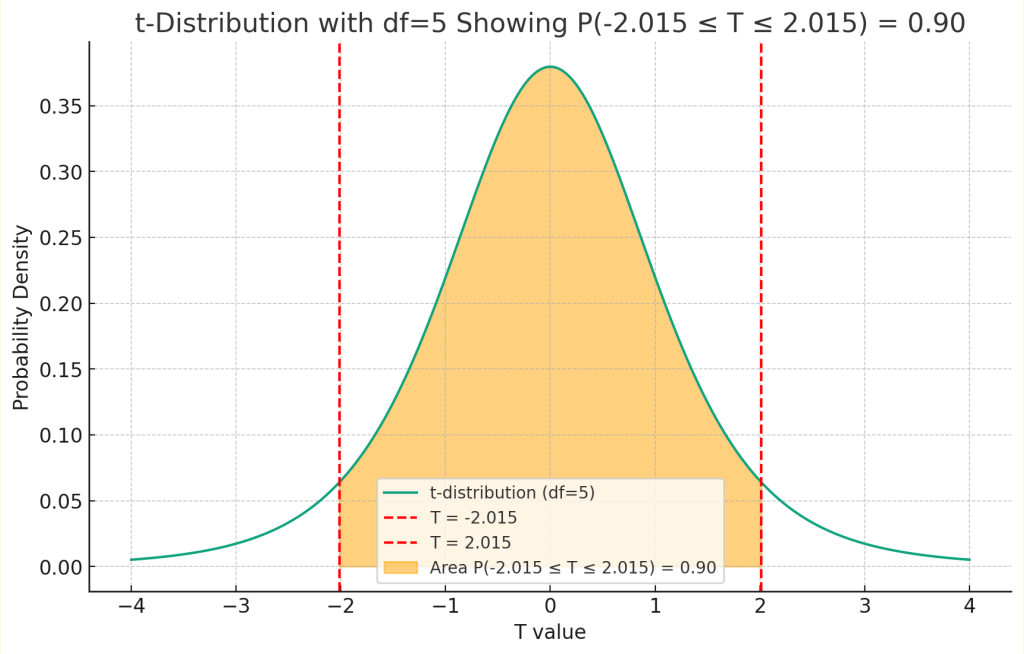


Figure 12: Graphic illustration of probability

2.2 F distribution

Definition(F distribution): Assume that random variables X and Y are independent of each other, $X \sim \chi^2_{(m)}, Y \sim \chi^2_{(n)}$, then it is said that $F = \frac{X/m}{Y/n}$ obeys the F distribution with degrees of freedom $(m,\ n)$, denoted as $F \sim F_{m,n}$, where m is called the first degree of freedom, n is called the second degree of freedom. The density function of the distribution is:

$$f(y) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\left(\frac{m}{n}\right)^{\frac{m}{2}} y^{\frac{m}{2}-1}\left(1+\frac{m}{n}y\right)^{-\frac{m+n}{2}}, \quad y > 0$$

The F-distribution can be illustrated as:

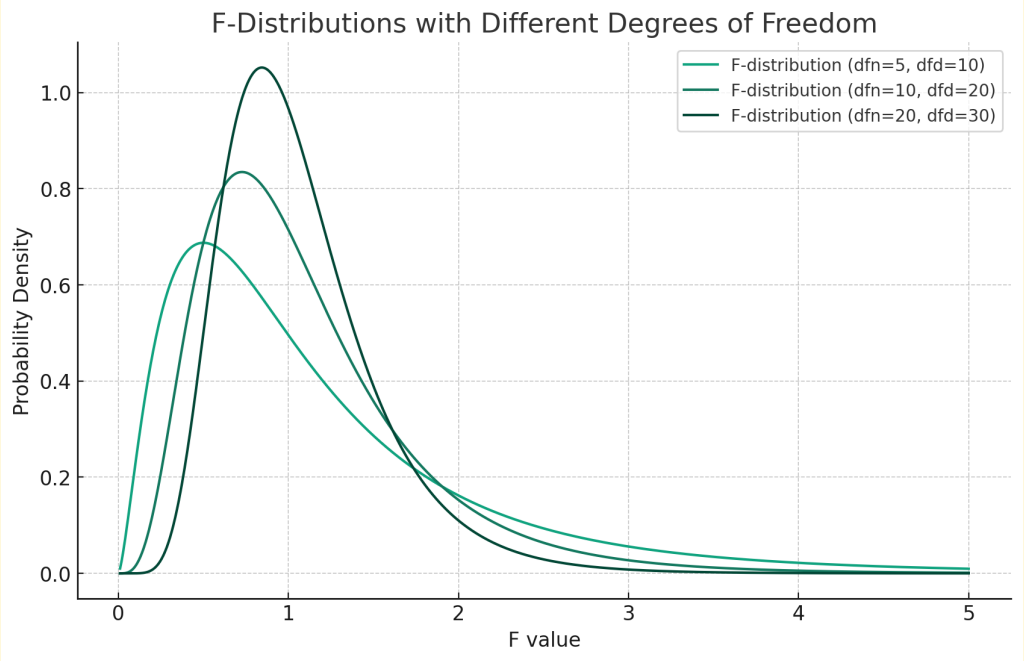


Figure 13: Different F distributions

Definition(Ratio of Independent R.Vs from Chi-squared)**:** Let U and V be two independent random variables from $X^2_{(m)}$ and $X^2_{(n)}$, respectively. Then

$$F = \frac{U/m}{V/n} \sim F_{m,n}$$

Example: Let $X_1, \ldots, X_m$ are iid (random sample) from $N(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_n$ are iid (random sample) from $N(\mu_2, \sigma_2^2)$, further the two samples are independent, then

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\left[(m-1)S_1^2/\sigma_1^2\right]/(m-1)}{\left[(n-1)S_2^2/\sigma_2^2\right]/(n-1)} = \frac{U/(m-1)}{V/(n-1)} \sim F_{m-1,n-1}$$

The α quantile of the $F_{m,n}$ distribution is denoted as $F_{\alpha,m,n}$, which means: for a given $\alpha(0 < \alpha < 1)$, When $F \sim F_{m,n}$, there is $P(F \leqslant F_{\alpha,m,n}) = \alpha$. The value $F_{\alpha,m,n}$ can be accessed using the F distribution table.

Theorem: For a F distribution $F_{m,n}$, we have $F_{\alpha,m,n} = \frac{1}{F_{1-\alpha,m,n}}$.

Proof. Assume that random variables X and Y are independent of each other, $X \sim \chi^2_{(m)}, Y \sim \chi^2_{(n)}$. Then:

$$F = \frac{X/m}{Y/n} \sim F_{m,n}, -\frac{1}{F} = \frac{Y/n}{X/m} \sim F_{n,m}$$

$$P(F > F_{\alpha,m,n}) = \alpha \text{ implies } P\left(\frac{1}{F} < \frac{1}{F_{\alpha,m,n}}\right) = \alpha.$$

Since $\frac{1}{F}$ has an $F_{n,m}$ distribution, for $\frac{1}{F}$ to have a lower tail probability of α, it must be less than the $1-\alpha$ upper tail critical value of the $F_{n,m}$ distribution, which is $F_{1-\alpha,n,m}$.
Therefore, $\frac{1}{F_{\alpha,m,n}} = F_{1-\alpha,n,m}$, or equivalently, $F_{\alpha,m,n} = \frac{1}{F_{1-\alpha,n,m}}$. ■

Example: If we take independent samples of size $m = 6$ and $n = 10$ from two normal populations with equal population variances, find the number b such that $P\left(\frac{S_1^2}{S_2^2} \leq b\right) = 0.95$.

Solution: solution: Because $m = 6$ and $n = 10$, and the variances are equal, then $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2}$ and has an F distribution with $m-1 = 5$ and $n-1 = 9$ degrees of freedom. From the F table, upper-tail area of 0.05 is 3.48 . i.e

$$P\left(\frac{S_1^2}{S_2^2} \geq 3.48\right) = 0.05$$

Hence $b = 3.48$

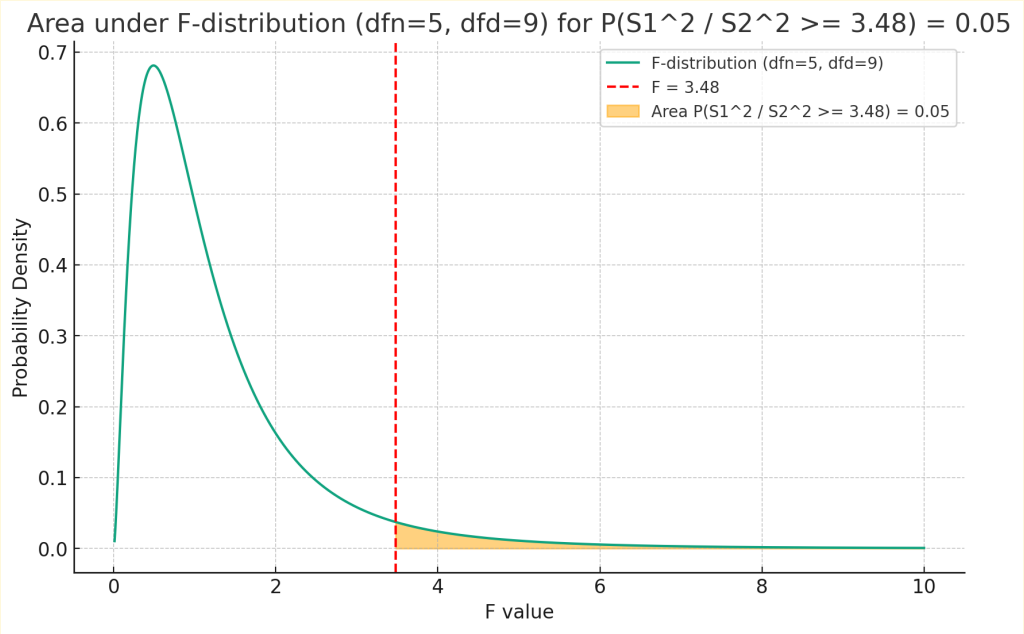


Figure 14: illustration of the area calcualted

3 Likelihood inference

Likelihood inference is a fundamental approach in statistical inference that focuses on utilizing the likelihood function as the primary tool for making inferences about a parameter or a set of parameters in a statistical model. This method is based on the data observed and the specified model, with minimal assumptions beyond those inherent in the model itself.

3.1 Likelihood function

Definition(Likelihood function): Let $\mathbf{x} = (X_1, \ldots, X_n)$ be iid sample of size n from probability mass function $p_X(t \mid \theta)$ (if X is discrete), or from density $f_X(t \mid \theta)$ (if X is continuous), where θ is a parameter (or vector of parameters). We define the likelihood of $\mathbf{x}$ given θ to be the "probability" of observing $\mathbf{x}$ if the true parameter is θ. If X is discrete,

$$L(\mathbf{x} \mid \theta) = p_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} p_X(x_i \mid \theta)$$

If X is continuous,

$$L(\mathbf{x} \mid \theta) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f_X(x_i \mid \theta)$$

Remark:

- In the Rosenthal book, $s = x_1, \ldots, x_n$ is sometimes used to denote data.
- $L(\cdot \mid s)$ is the likelihood function.
- $L(\theta \mid s)$ is the likelihood function at the value of θ and $L(\theta)$ is often used as shorthand notation.
- Constants that do not depend on θ are often dropped from the likelihood

Example: Suppose $\mathbf{x} = (x_1, x_2, x_3) = (1, 0, 1)$ are iid samples from $\mathrm{Ber}(\theta)$.

$$L(\mathbf{x} \mid \theta) = \prod_{i=1}^{3} p_X(x_i \mid \theta) = p_X(1 \mid \theta) \cdot p_X(0 \mid \theta) \cdot p_X(1 \mid \theta) = \theta(1-\theta)\theta = \theta^2(1-\theta)$$

Example: Suppose $\mathbf{x} = (x_1, x_2, x_3, x_4) = (3, 0, 2, 7)$ are iid samples from $\mathrm{Poi}(\theta)$. The samples mean we observed 3 events in the first unit of time, then 0 in the second, then 2 in the third, then 7 in the fourth. The

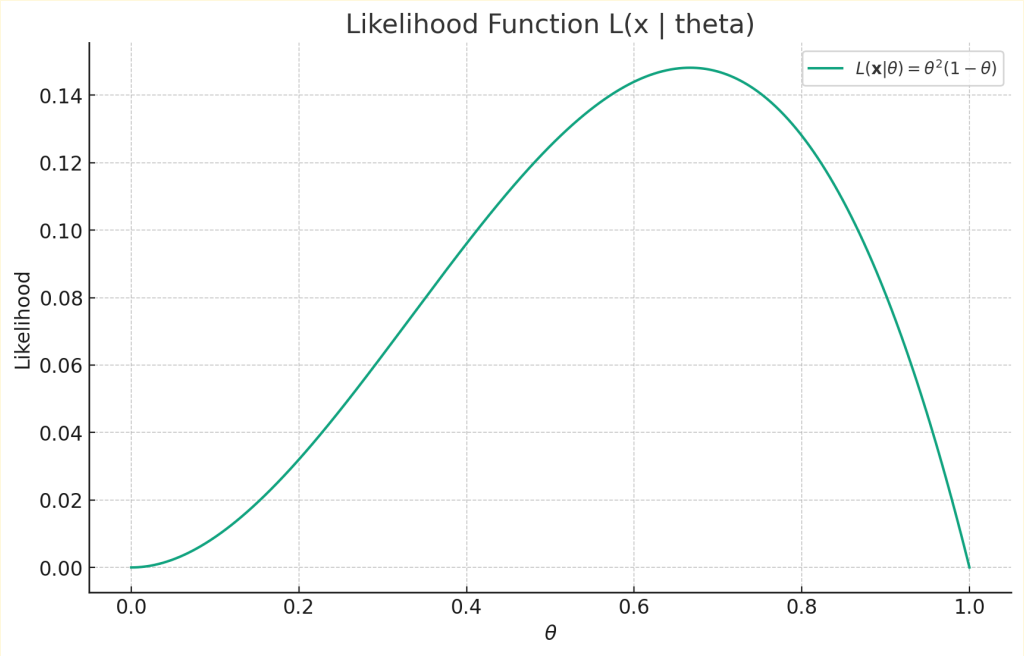


Figure 15: The graph of the function

likelihood is the "probability" of observing the data (just multiplying Poisson PMFs $p_X(k \mid \lambda) = e^{-\lambda}\frac{\lambda^k}{k!}$).

$$L(\mathbf{x} \mid \theta) = \prod_{i=1}^{4} p_X(x_i \mid \theta) = p_X(3 \mid \theta) \cdot p_X(0 \mid \theta) \cdot p_X(2 \mid \theta) \cdot p_X(7 \mid \theta) = \left(e^{-\theta}\frac{\theta^3}{3!}\right)\left(e^{-\theta}\frac{\theta^0}{0!}\right)\left(e^{-\theta}\frac{\theta^2}{2!}\right)\left(e^{-\theta}\frac{\theta^7}{7!}\right)$$

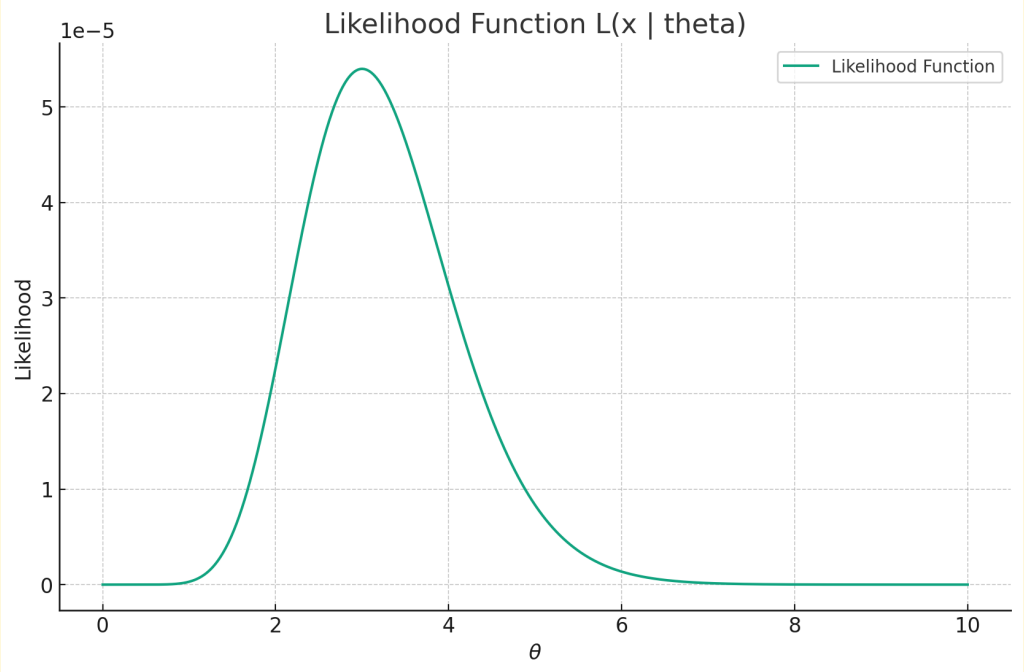


Figure 16: Graph of the function

Example: Suppose $\mathbf{x} = (x_1, x_2, x_3) = (3.22, 1.81, 2.47)$ are iid samples from $\mathrm{Exp}(\theta)$.
The samples mean we waited until three events happened (x_1, x_2, x_3), and it took 3.22 units of time until the first event, 1.81 until the second, and 2.47 until the third. The likelihood is the "probability" of observing the data. The likelihood is the "probability" of observing the data (just multiplying Exponential PDFs $f_X(y \mid \lambda) = \lambda e^{-\lambda y}$).

$$L(\mathbf{x} \mid \theta) = \prod_{i=1}^{3} f_X(x_i \mid \theta) = f_X(x_1 \mid \theta) \cdot f_X(x_2 \mid \theta) \cdot f_X(x_3 \mid \theta) = \left(\theta e^{-3.22\theta}\right)\left(\theta e^{-1.81\theta}\right)\left(\theta e^{2.47\theta}\right)$$

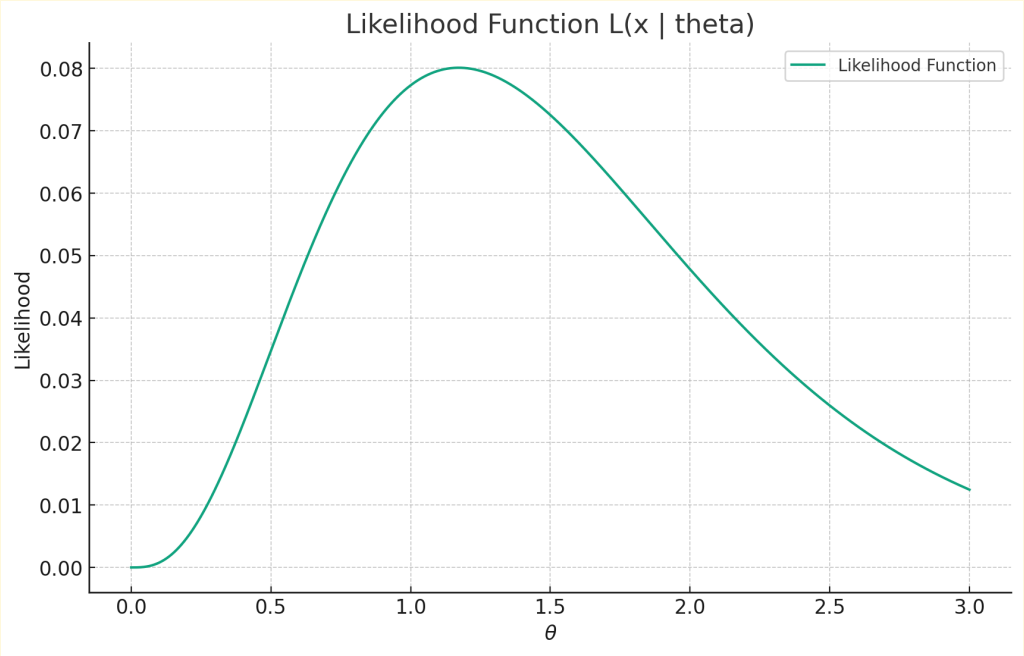


Figure 17: Graph of the function

The likelihood principle

If two model and data combinations yield equivalent likelihood functions, then inferences about the unknown parameter must be the same.

Example: Imagine two experiments designed to estimate the probability θ of getting a head when flipping a biased coin, where θ is the unknown parameter of interest.

- **Experiment 1:** A coin is flipped 10 times, resulting in 7 heads and 3 tails.
- **Experiment 2:** A coin is flipped 100 times, resulting in 70 heads and 30 tails.

Assuming each coin flip is independent, the likelihood function for each experiment, given the data, can be modeled using the binomial distribution:

- For Experiment 1, the likelihood function is $L\left(\theta \mid \text{data}_1\right) = \left(\begin{array}{c} 10 \\ 7 \end{array} \right) \theta^7 (1-\theta)^3$.
- For Experiment 2, the likelihood function is $L\left(\theta \mid \text{data}_2\right) = \left(\begin{array}{c} 100 \\ 70 \end{array} \right) \theta^{70} (1-\theta)^{30}$.

According to the likelihood principle, since the likelihood functions from both experiments are proportional to each other when ignoring the constants (which are irrelevant for inference about θ), any inference about θ should be the same regardless of whether we are looking at the data from Experiment 1 or Experiment 2. This means that our estimate of θ, confidence intervals, or any other inferential statistics about θ should be consistent across both experiments.

3.2 Sufficient statistics

The likelihood function, $L(\cdot \mid s)$, quantifies how likely it is to observe a given set of data, s, under various parameter values of a statistical model. The "." symbol here represents the parameter or parameters for which the likelihood is being calculated.

If the likelihood function for one set of data, s_1, is equal to a constant positive multiple, c, of the likelihood function for another set of data, s_2, this means mathematically that $L\left(\cdot \mid s_1\right) = cL\left(\cdot \mid s_2\right)$, where $c > 0$.

This mathematical relationship implies that both data sets, s_1 and s_2, provide the same information about the parameters of the model, in terms of likelihood. The constant c does not affect the shape or the peak of the likelihood function relative to the parameters; it only changes the scale. Since inferential conclusions (such

as estimates of model parameters, confidence intervals, or hypothesis tests) are based on the shape and peak (or peaks) of the likelihood function, not its scale, s_1 and s_2 are considered equivalent for the purpose of making inferences about the model's parameters.

Thus, regardless of whether the data observed were s_1 or s_2, the conclusions drawn about the model's parameters would be the same. Thus, either data set (or equivalently, the statistic that encapsulates this data) is sufficient for making inferences about the model parameters. And this leads to the formal definition of sufficient statistics:

Definition(Sufficient statistics): A function T defined on the sample space S is called a sufficient statistic for the model if, whenever $T(s_1) = T(s_2)$, then

$$L(\cdot \mid s_1) = c(s_1, s_2) L(\cdot \mid s_2)$$

for some constant $c(s_1, s_2) > 0$.

Example: Suppose that $S = \{1, 2, 3, 4\}, \Omega = \{a, b\}$, and the two probability distributions are given by the following table.

	$s=1$	$s=2$	$s=3$	$s=4$
$\theta = a$	$1/2$	$1/6$	$1/6$	$1/6$
$\theta = b$	$1/4$	$1/4$	$1/4$	$1/4$

The table provides the probabilities of observing each data value under the two different parameter settings ($\theta = a$ and $\theta = b$):

- For $\theta = a$: $P(S=1) = \frac{1}{2}, P(S=2) = P(S=3) = P(S=4) = \frac{1}{6}$.
- For $\theta = b$: $P(S=1) = \frac{1}{4}, P(S=2) = P(S=3) = P(S=4) = \frac{1}{4}$.

The likelihood of observing each of $s = 2, 3, 4$ given θ is the same, regardless of whether $\theta = a$ or $\theta = b$. This means that observing any of these values provides equivalent information about θ, as they result in the same likelihood ratios $(L(a \mid s)/L(b \mid s)$) for $s = 2, 3, 4$.
The function $T : S \to \{0, 1\}$ is defined such that $T(1) = 0$ and $T(2) = T(3) = T(4) = 1$. This transformation reduces the original sample space from four possible outcomes to two, based on the equivalence in the likelihood information provided by $s = 2, 3, 4$. The value 0 corresponds to observing $s = 1$, and the value 1 corresponds to observing any of $s = 2, 3, 4$.
The statistic T is sufficient for θ because it captures all the information in the sample about which distribution ($\theta = a$ or $\theta = b$) is more likely. The fact that $L(\cdot \mid 2) = L(\cdot \mid 3) = L(\cdot \mid 4)$ means that from the perspective of inference about θ, knowing whether $T = 0$ or $T = 1$ is as informative as knowing the specific value of s from $\{2, 3, 4\}$. Thus, T simplifies the model without losing information about θ, which is the essence of a sufficient statistic: it retains all the information about the parameter of interest while potentially reducing the complexity of the data.

And we have a easier way to identify sufficient statistics:

Factorization theorem

Theorem: Let T be a statistic of the random sample $X_1, X_2, \ldots, X_n$.
T is a sufficient statistic for a parameter θ if and only if the likelihood can be factored into two non-negative functions,

$$L(\theta) = g(t, \theta) \times h(x_1, \ldots, x_n)$$

where $g(t, \theta)$ is a function of t and θ and $h(x_1, \ldots, x_n)$ is not a function of θ.

Remark:

- Sufficient statistic always exists and it is not unique.
- Any one-to-one function of a sufficient statistic is a sufficient statistic.
- A random sample itself is a sufficient statistic.

Proof. Now, we will prove the factorization theorem in both directions:

- "$\Longrightarrow$". Let $t = T(x_1, \ldots, x_n)$ be the observed value of the statistic for the sample $(x_1, \ldots, x_n)$. Since T is sufficient, $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \mid T = t)$ is independent of θ and therefore the likelihood can be factorized as

$$\begin{aligned} L(\theta; x_1, \ldots, x_n) &= \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n; \theta) \\ &= \mathbb{P}(\{X_1 = x_1, \ldots, X_n = x_n\} \cap \{T = t\}; \theta) \\ &= p(T = t; \theta)\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \mid T = t), \end{aligned}$$

  which agrees with the desired factorization just by taking

$$g(t, \theta) = p(T = t; \theta), \quad h(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \mid T = t).$$

  Therefore,

$$L(\theta; x_1, \ldots, x_n) = g(t, \theta)h(x_1, \ldots, x_n).$$

- " $\Longleftarrow$ ". Assume now that the factorization

$$L(\theta; x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n; \theta) = g(t, \theta)h(x_1, \ldots, x_n)$$

  We define the set

$$A_t = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : T(x_1, \ldots, x_n) = t\}.$$

  Then,

$$\begin{aligned} p(T = t; \theta) &= \sum_{(x_1, \ldots, x_n) \in A_t} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n; \theta) \\ &= g(t, \theta) \sum_{(x_1, \ldots, x_n) \in A_t} h(x_1, \ldots, x_n), \end{aligned}$$

  so therefore,

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \mid T = t; \theta) = \begin{cases} \frac{h(x_1, \ldots, x_n)}{\sum_{(x_1, \ldots, x_n) \in A_t} h(x_1, \ldots, x_n)} & \text{if } T(x_1, \ldots, x_n) = t \\ 0 & \text{if } T(x_1, \ldots, x_n) \neq t. \end{cases}$$

  Since $h(x_1, \ldots, x_n)$ does not depend on θ, then the conditional distribution of $(X_1, \ldots, X_n)$ given T does not depend on θ. Therefore, T is sufficient.

■

Example: Let $X_1, X_2, \ldots, X_n$ be a random sample from $\mathrm{Exp}(\lambda)$. Show that $\bar{X}$ is a sufficient statistic for parameter λ.

Proof. The pdf of an exponential distribution with rate parameter λ is given by:

$$f(x \mid \lambda) = \lambda e^{-\lambda x}, \quad x \geq 0, \lambda > 0$$

Given a random sample $X_1, X_2, \ldots, X_n$ from $\mathrm{Exp}(\lambda)$, the joint likelihood function of the sample is the product of the individual densities (since the observations are independent):

$$L(\lambda; X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

According to the Factorization Theorem, $\bar{X}$ will be a sufficient statistic for λ if we can express the likelihood function as a product of two functions, where one function depends on $\bar{X}$ and λ, and the other does not depend on λ.
Notice that in the likelihood function above, we can rewrite $\sum_{i=1}^{n} x_i$ as $n\bar{X}$, where $\bar{X}$ is the sample mean $\frac{1}{n}\sum_{i=1}^{n} x_i$. Therefore, the likelihood function can be rewritten as:

$$L(\lambda; X_1, X_2, \ldots, X_n) = \lambda^n e^{-\lambda n \bar{X}}$$

Here we can see that $g(T, \lambda) = \lambda^n e^{-\lambda n \bar{X}}$, where $T = \bar{X}$, is a function of both $\bar{X}$ and λ. And $h(x_1, x_2, \ldots, x_n) = 1$ ■

Example: Suppose $x_1 \ldots, x_n \overset{iid}{\sim} \text{Gamm}(\alpha = 2, \beta)$

$$f(x, \alpha = 2, \beta) = \frac{1}{\beta^2} x e^{-x/\beta}, \beta, x > 0$$

Find a sufficient statistics T for β.
Solution: To find a sufficient statistic T for β using the factorization theorem, we start with the given probability density function (pdf) for a Gamma distribution with parameters $\alpha = 2$ and β, which is given as:

$$f(x; \alpha = 2, \beta) = \frac{1}{\beta^2} x e^{-x/\beta}, \quad \beta, x > 0$$

Given $x_1, \ldots, x_n \overset{iid}{\sim} \text{Gamma}(\alpha = 2, \beta)$, the joint pdf of the sample $X_1, X_2, \ldots, X_n$ is the product of the individual pdfs because the observations are independent and identically distributed (iid). Therefore, the joint pdf is:

$$L(\beta; x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\beta^2} x_i e^{-x_i/\beta} = \frac{1}{\beta^{2n}} \left(\prod_{i=1}^{n} x_i \right) e^{-\sum_{i=1}^{n} x_i/\beta}$$

To apply the factorization theorem, we need to express this likelihood in the form $g(t, \beta) \times h(x_1, \ldots, x_n)$, where g is a function of the statistic t and the parameter β, and h does not depend on β.
From the joint pdf, we can see that:

$$L(\beta; x_1, \ldots, x_n) = \frac{1}{\beta^{2n}} e^{-\sum_{i=1}^{n} x_i/\beta} \times \left(\prod_{i=1}^{n} x_i \right)$$

Here, $g(t, \beta) = \frac{1}{\beta^{2n}} e^{-\sum_{i=1}^{n} x_i/\beta}$ and $h(x_1, \ldots, x_n) = \prod_{i=1}^{n} x_i$. The function h does not depend on β, fulfilling the factorization theorem's condition.
The term that involves β is:

$$g(t, \beta) = \frac{1}{\beta^{2n}} e^{-\sum_{i=1}^{n} x_i/\beta}$$

This suggests that the sufficient statistic T for β is related to the components of $g(t, \beta)$. Specifically, the sum $\sum_{i=1}^{n} x_i$ appears in the exponent that involves β, and the power of β in the denominator suggests that the sample size n also plays a role. However, for the purpose of sufficiency and the factorization theorem, the relevant part is just the sum of the observations:

$$T = \sum_{i=1}^{n} x_i$$

This sum T is a sufficient statistic for β because the likelihood can be factored into a product where one factor, $g(t, \beta)$, depends only on T and β, and the other factor, $h(x_1, \ldots, x_n)$, does not depend on β. This satisfies the conditions of the factorization theorem, confirming that T is indeed sufficient for β.
Example: Let $X_1, X_2, \ldots, X_n$ be a random sample from a uniform distribution on $(0, \theta)$.
Find a sufficient statistic for θ.
Solution: To find a sufficient statistic for θ when $X_1, X_2, \ldots, X_n$ are drawn from a uniform distribution on $(0, \theta)$, we start by writing the probability density function (pdf) for a single observation X_i :

$$f(x_i; \theta) = \frac{1}{\theta}, \quad 0 < x_i < \theta$$

Given that the sample $X_1, X_2, \ldots, X_n$ is independent and identically distributed (iid), the joint pdf of the sample is the product of the individual pdfs:

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\theta} = \frac{1}{\theta^n}, \quad 0 < x_i < \theta \text{ for all } i$$

However, for the joint pdf to be non-zero, the condition that all x_i are less than θ must hold. This means that the maximum of all x_i, denoted as $\max(x_1, x_2, \ldots, x_n)$, must also be less than or equal to θ. Hence, the support of the likelihood function depends on θ, and we adjust our expression for the likelihood to include this condition:

$$L(\theta; x_1, \ldots, x_n) = \frac{1}{\theta^n}, \quad \text{for } \max(x_1, x_2, \ldots, x_n) \leq \theta$$

The likelihood function can be written, incorporating this condition, as:

$$L(\theta; x_1, \ldots, x_n) = \begin{cases} \frac{1}{\theta^n} & \text{if } \theta \geq \max(x_1, x_2, \ldots, x_n) \\ 0 & \text{otherwise} \end{cases}$$

To apply the factorization theorem and identify a sufficient statistic for θ, we note that the likelihood function can be factorized into two parts: one that depends on θ and the sample (through $\max(x_1, x_2, \ldots, x_n)$) and one that does not depend on θ. However, in this case, the entire dependence on θ and the sample is through the term $\frac{1}{\theta^n}$ and the condition $\theta \geq \max(x_1, x_2, \ldots, x_n)$.
Thus, the sufficient statistic for θ is $T(X_1, X_2, \ldots, X_n) = \max(X_1, X_2, \ldots, X_n)$, as the value of θ primarily affects the likelihood through its relationship to the maximum value of the sample. This meets the factorization theorem's criteria, showing that the maximum value of the sample is a sufficient statistic for θ in the case of a uniform distribution on $(0, \theta)$.

Among all possible sufficient statistics, a minimum sufficient statistic is the one that provides the most compressed or reduced form of the data without losing any information about parameter θ.

Application of the factorization theorem typically leads to a minimal sufficient statistic that provides the best summary of the information in the data. But factorization theorem does not guarantee a minimum sufficient statistic.

Definition(Minimum sufficient statistic)**:** A sufficient statistic T is minimal sufficient if it is a function of every other sufficient statistic.

3.3 Maximum likelihood estimation

Definition(Maximum likelihood estimate)**:** The maximum likelihood estimate (MLE), denoted by $\hat{\theta}$, is the value of θ that maximizes $L(\theta)$.

Intuitively, the MLE is just the value of θ which maximizes the "probability" of seeing the data $L(\mathbf{x} \mid \theta)$.

Procedures of calculating MLE

1. Suppose $X_1, \ldots, X_n \sim^{\text{iid}} f(x; \theta)$.
   Write down the likelihood function as
   $$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$
   If the the likelihood function is differentiable with respect to θ, follow steps $2 - 4$.
2. Take the natural log of the likelihood
   $$\ell(\theta) = \log L(\theta)$$
   and collect terms involving θ.
3. Find the value of $\theta \in \Theta$, denoted $\hat{\theta}$, that maximizes $\ell(\theta) = \log L(\theta)$ use calculus.
4. Check that θ obtained in Step 3 is a unique maximum by inspecting the second derivative of $\ell(\theta) = \log L(\theta)$ with respect to θ. If
   $$\frac{d^2\ell(\theta)}{d\theta^2} < 0$$
   at $\theta = \hat{\theta}$, then $\hat{\theta}$ is confirmed as the MLE of θ.

The log likelihood and likehood function and intersection of maximum point can be illustrated as:

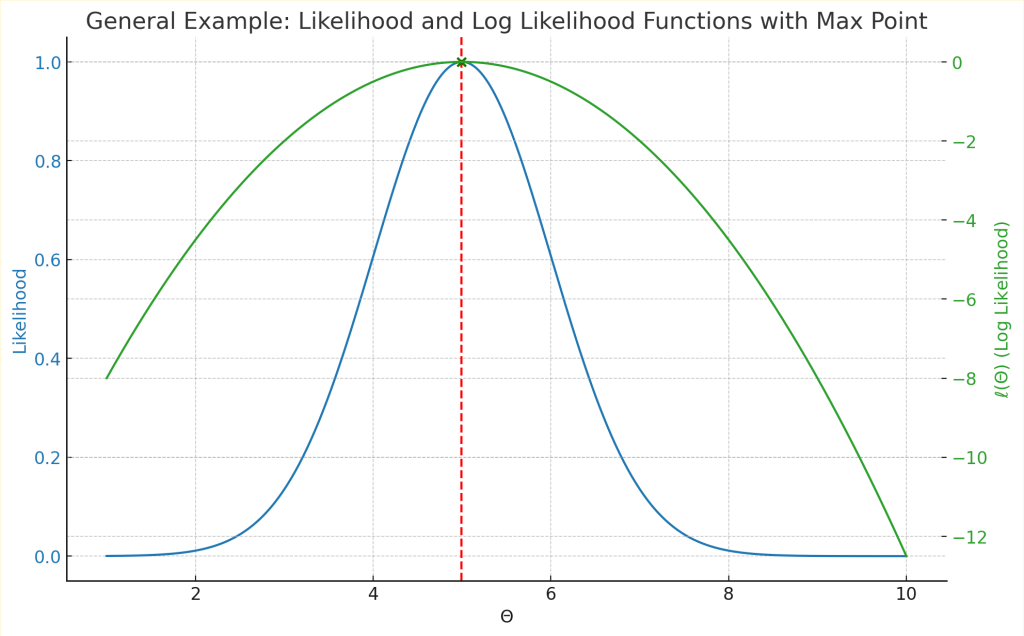


Figure 18: intersection of two likelihood functions

Elementary properties of MLE

- Sufficiency property: If T is sufficient for θ, then the maximum likelihood estimator of θ is a function of T.
- Invarience property: If $g(\theta)$ is a one-to-one function of θ and if $\hat{\theta}$ is the MLE for θ, then the MLE of $g(\theta)$ is $g(\hat{\theta})$

Example: Consider a random sample of size n from a Poisson distribution with parameter $\lambda > 0$.
Find the MLE of λ and λ^2 analytically.
Solution: For a Poisson distribution, the probability mass function (pmf) for a single observation X_i is:

$$f\left(x_i;\lambda\right)=\frac{e^{-\lambda}\lambda^{x_i}}{x_i!},\quad x_i=0,1,2,\ldots$$

Given $X_1,\ldots,X_n\sim$ iid Poisson(λ), the likelihood function is:

$$L(\lambda)=\prod_{i=1}^{n}\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}=e^{-n\lambda}\lambda^{\sum_{i=1}^{n}x_i}\prod_{i=1}^{n}\frac{1}{x_i!}$$

Taking the natural logarithm of the likelihood function, we get the log-likelihood function:

$$\ell(\lambda)=\log L(\lambda)=-n\lambda+\left(\sum_{i=1}^{n}x_i\right)\log\lambda-\sum_{i=1}^{n}\log x_i!$$

To find the MLE of λ, we differentiate the log-likelihood function with respect to λ and set the derivative equal to zero:

$$\frac{d\ell(\lambda)}{d\lambda}=-n+\frac{\sum_{i=1}^{n}x_i}{\lambda}$$

Setting this equal to zero to find the critical points:

$$-n+\frac{\sum_{i=1}^{n}x_i}{\lambda}=0\Longrightarrow\lambda=\frac{\sum_{i=1}^{n}x_i}{n}$$

Thus, the MLE of λ, denoted $\hat{\lambda}$, is:

$$\hat{\lambda}=\frac{\sum_{i=1}^{n}x_i}{n}$$

To verify that this critical point is a maximum, we check the second derivative of the log-likelihood:

$$\frac{d^2\ell(\lambda)}{d\lambda^2} = -\frac{\sum_{i=1}^n x_i}{\lambda^2}$$

Substituting $\hat{\lambda}$, we find:

$$\frac{d^2\ell(\hat{\lambda})}{d\lambda^2} = -\frac{n}{\hat{\lambda}^2} < 0$$

This confirms that $\hat{\lambda}$ is indeed the MLE of λ.
To find the MLE of λ^2, we use the invariance property of MLEs, which states that if $\hat{\theta}$ is the MLE of θ, then for any function g, the MLE of $g(\theta)$ is $g(\hat{\theta})$.
Therefore, the MLE of λ^2 is:

$$\widehat{\lambda^2} = \left(\frac{\sum_{i=1}^n x_i}{n}\right)^2$$

Example: Let $X_1, \ldots, X_n$ be random sample from $U(0, \theta)$. Find the MLE of θ.
Solution: To compute the likelihood, we first need the individual density functions. Recall

$$f_X(x;\theta) = \begin{cases} \frac{1}{\theta} & 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases}$$

Let's actually define an indicator function for whether or not some boolean condition A is true or false:

$$I_A = \begin{cases} 1 & A \text{ is true} \\ 0 & A \text{ is false} \end{cases}$$

This way, we can rewrite the uniform density in one line as ($1/\theta$ for $0 \le x \le \theta$ and 0 otherwise):

$$f_X(x;\theta) = \frac{1}{\theta} I_{\{0 \le x \le \theta\}}$$

And our likelihood is

$$L(\theta; x_1, x_2, \ldots, x_n) = \prod_{i=1}^n f_X(x_i;\theta) = \prod_{i=1}^n \frac{1}{\theta} I_{\{0 \le x_i \le \theta\}} = \frac{1}{\theta^n} I_{\{0 \le x_1, \ldots, x_n \le \theta\}}$$

Above is a graph of $\frac{1}{\theta^n}$ with the indicator function, and so if we wanted to maximize this function, we should

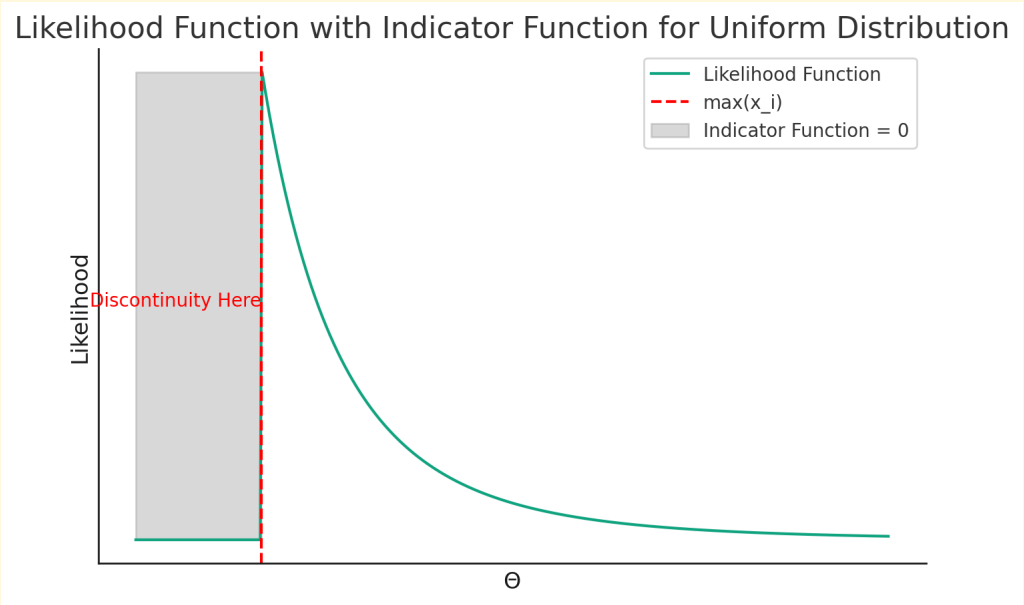


Figure 19: Illustration of the likelihood

choose $\theta = 0$. But remember that the likelihood, was $\frac{1}{\theta^n} I_{\{0 \le x_1, \ldots, x_n \le \theta\}}$, which can also be written as $\frac{1}{\theta^n} I_{\{x_{\max} \le \theta\}}$,

because all the samples are $\leq \theta$ if and only if the maximum is.
We have $x_{\max} \leq \theta$, but zeroed it out otherwise. So now we can see that our maximum likelihood estimator should be $\hat{\theta}_{MLE} = x_{\max} = \max\{x_1, x_2, \ldots, x_n\}$, since it achieves the highest value.
Remember $x_1, \ldots, x_n \sim \text{Unif}(0, \theta)$, so θ has to be at least as large as the biggest x_i, because if it's not as large as the biggest x_i, then it would have been impossible for that uniform to produce that largest x_i. For example, if our samples were $x_1 = 2.53, x_2 = 8.55, x_3 = 4.12$, our θ had to be at least 8.55 (the maximum sample), because if it were 7 for example, then $\text{Unif}(0, 7)$ could not possibly generate the sample 8.55.
So our likelihood remember $\frac{1}{\theta^n}$ would have preferred as small a θ as possible to maximize it, but subject to $\theta \geq x_{\max}$. Therefore the "compromise" was reached by making them equal.

3.4 Method of moments estimator

The idea behind Method of Moments estimation (MME) is that: to find a good estimator, we should have the true and sample moments match as best we can. That is, I should choose the parameter θ such that the first true moment $\mathbb{E}[X]$ is equal to the first sample moment $\bar{x}$.

Definition(Method of moments estimation)**:** Let $\mathbf{x} = (x_1, \ldots, x_n)$ be iid realizations (samples) from probability mass function $p_X(t; \theta)$ (if X is discrete), or from density $f_X(t; \theta)$ (if X continuous), where θ is a parameter (or vector of parameters).
We then define the method of moments estimator $\hat{\theta}_{MoM}$ of $\theta = (\theta_1, \ldots, \theta_k)$ to be a solution (if it exists) to the k simultaneous equations where, for $j = 1, \ldots, k$, we set the j^{th} (true) moment equal to the j^{th} sample moment:

$$\mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$\ldots$$
$$\mathbb{E}\left[X^k\right] = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

Example: Let's say $x_1, x_2, \ldots, x_n$ are iid samples from $X \sim \text{Exp}(\theta)$. (These values might look like $x_1 = 3.21, x_2 = 5.11, x_3 = 4.33$, etc.) What is the MME estimator of θ?
Solution: Solution We have $k = 1$ (since only one parameter). We then set the first true moment to the first sample moment as follows (recall that $\mathbb{E}[\text{Exp}(\lambda)] = \frac{1}{\lambda}$):

$$\mathbb{E}[X] = \frac{1}{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Solving for θ (just taking inverse), we get:

$$\hat{\theta}_{MoM} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i}$$

Notice that in this case, the MME estimator agrees with the MLE (Maximum Likelihood Estimator)

$$\hat{\theta}_{MoM} = \hat{\theta}_{MLE} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i}$$

Example: Let's say $x_1, x_2, \ldots, x_n$ are iid samples from $X \sim \text{Poi}(\theta)$. (These values might look like $x_1 = 13, x_2 = 5, x_3 = 4$, etc.) What is the MME estimator of θ ?
Solution: We have $k = 1$ (since only one parameter). We then set the first true moment to the first sample moment as follows (recall that $\mathbb{E}[\text{Poi}(\lambda)] = \lambda$) :

$$\mathbb{E}[X] = \theta = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Solving for θ, we get:

$$\hat{\theta}_{MoM} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

In this case, again, the MME estimator agrees with the MLE.
Example: Let's say $x_1, x_2, \ldots, x_n$ are iid samples from $X \sim \mathcal{N}(\theta_1, \theta_2)$. (These values might look like $x_1 = -2.321, x_2 = 1.112, x_3 = -5.221$, etc.) What is the MME estimator of the vector $\theta = (\theta_1, \theta_2)$ (θ_1 is the mean, and θ_2 is the variance)?
Solution: We have $k = 2$ (since now we have two parameters $\theta_1 = \mu$ and $\theta_2 = \sigma^2$). Notice $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, so rearranging we get $\mathbb{E}[X^2] = \mathrm{Var}(X) + \mathbb{E}[X]^2$. Let's solve for θ_1 first:
Again, we set the first true moment to the first sample moment:

$$\mathbb{E}[X] = \theta_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Solving for θ_1, we get:

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Now let's use our result for $\hat{\theta}_1$ to solve for $\hat{\theta}_2$ (recall that $\mathbb{E}[X^2] = \mathrm{Var}(X) + \mathbb{E}[X]^2 = \theta_2 + \theta_1^2$)

$$\mathbb{E}\left[X^2\right] = \theta_2 + \theta_1^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

Solving for θ_2, and plugging in our result for $\hat{\theta}_1$, we get:

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$$

If we use MLE, we would have the same result.

3.5 Inferences based on the MLE

Definition(Mean square error)**:** The mean squared error of a point estimator $\hat{\theta}$ of θ is defined as

$$\mathrm{MSE}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right] = [b(\hat{\theta})]^2 + \mathrm{Var}(\hat{\theta})$$

where $b(\hat{\theta}) = E(\hat{\theta} - \theta)$.

Example: Consider a random sample $X_1, \ldots X_n$ arising from a Poisson distribution with mean λ. We showed that the maximum likelihood estimator (MLE) for μ is $\bar{X}$ in the previous example.
Find the MSE of $\bar{X}$.
Solution: For a Poisson distribution, the mean and variance are both equal to λ. Since $\bar{X}$ is an unbiased estimator of λ (i.e., $E(\bar{X}) = \lambda$), we have:

- $$b(\bar{X}) = E(\bar{X} - \lambda) = E(\bar{X}) - \lambda = \lambda - \lambda = 0$$

  The bias $b(\bar{X})$ is zero because $\bar{X}$ is an unbiased estimator of λ.
- The variance of the sample mean $\bar{X}$ for a Poisson distribution is:

$$\mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X)}{n} = \frac{\lambda}{n}$$

Using the formula for MSE:

$$\mathrm{MSE}(\bar{X}) = [b(\bar{X})]^2 + \mathrm{Var}(\bar{X}) = 0^2 + \frac{\lambda}{n} = \frac{\lambda}{n}$$

Therefore, the Mean Squared Error (MSE) of $\bar{X}$, the MLE for λ from a Poisson distribution, is $\frac{\lambda}{n}$.
Example: Suppers $x_1, \cdots, x_n \sim^{i.i.d} N(\mu, \sigma^2)$, where μ is known, σ^2 is unknown. Find the MLE of σ^2. Then Find the MSE of the MLE.

Solution: Given $X_1, \ldots, X_n \sim^{\text{i.i.d}} N(\mu, \sigma^2)$, the likelihood function of the sample is:

$$L\left(\sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

The log-likelihood function $\ell\left(\sigma^2\right)$ is:

$$\ell\left(\sigma^2\right) = \log L\left(\sigma^2\right) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Differentiating $\ell\left(\sigma^2\right)$ with respect to σ^2 and setting the derivative equal to zero gives the MLE of σ^2 :

$$\frac{d\ell\left(\sigma^2\right)}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\left(\sigma^2\right)^2} = 0$$

Solving for σ^2, we find:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

Thus, the MLE of σ^2 is:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

The mean squared error (MSE) of an estimator $\hat{\theta}$ is defined as:

$$\text{MSE}(\hat{\theta}) = E\left[(\hat{\theta}-\theta)^2\right] = [b(\hat{\theta})]^2 + \text{Var}(\hat{\theta})$$

For $\hat{\sigma}^2$, the bias $b\left(\hat{\sigma}^2\right) = E\left[\hat{\sigma}^2\right] - \sigma^2$. However, $\hat{\sigma}^2$ is an unbiased estimator of σ^2 in the normal distribution, so $b\left(\hat{\sigma}^2\right) = 0$. Thus, the MSE is simply the variance of $\hat{\sigma}^2$.
The variance of $\hat{\sigma}^2$ for a normal distribution is known to be:

$$\text{Var}\left(\hat{\sigma}^2\right) = \frac{2\sigma^4}{n}$$

Therefore, the MSE of $\hat{\sigma}^2$ is:

$$\text{MSE}\left(\hat{\sigma}^2\right) = \frac{2\sigma^4}{n}$$

Consistency and MSE

Theorem: Let $\hat{\theta}$ be an estimator of a parameter θ. If

$$\lim_{n\to\infty} \text{MSE}(\hat{\theta}) = 0$$

then $\hat{\theta} \to^p \theta$.

Proof. Given that $\lim_{n\to\infty} \text{MSE}(\hat{\theta}) = 0$, it implies both:

1. $\lim_{n\to\infty} \text{Var}(\hat{\theta}) = 0$
2. $\lim_{n\to\infty} [b(\hat{\theta})]^2 = 0$, which implies $\lim_{n\to\infty} b(\hat{\theta}) = 0$

From 1, since the variance of $\hat{\theta}$ goes to 0 , it means that $\hat{\theta}$ becomes increasingly concentrated around its mean $E[\hat{\theta}]$.
From 2 , since the bias goes to 0 , the mean $E[\hat{\theta}]$ converges to θ.
To connect these two points to convergence in probability, consider Chebyshev's inequality, which for any $\epsilon > 0$ gives:

$$P(|\hat{\theta} - E[\hat{\theta}]| > \epsilon) \leq \frac{\text{Var}(\hat{\theta})}{\epsilon^2}$$

As n approaches infinity, $\text{Var}(\hat{\theta}) \to 0$, making the right side of the inequality go to 0 , which implies:

$$\lim_{n\to\infty} P(|\hat{\theta} - E[\hat{\theta}]| > \epsilon) = 0$$

Given that $E[\hat{\theta}]$ converges to θ (from the bias going to 0), we can replace $E[\hat{\theta}]$ with θ in our probabilistic statement to get:

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| > \epsilon) = 0$$

Thus, $\hat{\theta} \to^p \theta$, proving that if the MSE of $\hat{\theta}$ goes to 0 as n approaches infinity, then $\hat{\theta}$ converges in probability to θ. ■

Definition(Confidence interval)**:** The $(1-\alpha)$ confidence interval (Cl) for a parameter θ is an interval $C_n = (a, b)$ where $a = a(X_1, \ldots, X_n)$ and $b = b(X_1, \ldots, X_n)$ such that:

$$P(a \leq \theta \leq b) \geq 1 - \alpha, \text{ for all } \theta \in \boldsymbol{\theta}$$

$1 - \alpha$ is called the confidence level of the interval.

Remark: We also often referred then as a two-sided confidence interval or a $(1-\alpha) \times 100\%$ confidence interval.
Remark: Because of the discrete nature of the probability mass function (PMF), it may not always be possible to construct a confidence interval that exactly corresponds to the desired confidence level $(1 - \alpha)$.
Remark: Since the interval can be adjusted infinitesimally due to the nature of continuous data, it's typically possible to construct a confidence interval that achieves the exact desired confidence level $(1 - \alpha)$.
Example: As a example, suppose we have a 95% confidence interval, and we can intepretate the meaning in the following way:

- **Same experiment repeated:** In this context, saying that "95% of the intervals we construct will contain the true parameter value" means that if we were to repeat this experiment an infinite number of times, constructing a 95% CI for θ each time, then we would expect 95% of these intervals to actually contain the true value of θ. This does not mean that any given interval has a 95% chance of containing θ; rather, it is about the long-term proportion of such intervals that would contain θ if the experiment were repeated many times under the same conditions.
- **Different experiments for different parameters:** Here, the statement that "95% of the intervals we construct will contain the true parameter value" extends the first scenario's logic to a broader context. It implies that if we conduct various experiments, each aimed at estimating different parameters, and for each experiment, we construct a 95% CI for the parameter of interest, then over the long run, 95% of these confidence intervals will contain the true values of their respective parameters. This illustrates the general applicability of the confidence level concept across different studies and parameters, not just the repetition of the same experiment.

We have a general strategy to find the bounded quantity for the confidence interval. Suppose that we are interested in finding two values x_h and x_l such that

$$P(x_l \leq X \leq x_h) = 1 - \alpha.$$

One way to do this, is to chose x_l and x_h such that

$$P(X \leq x_l) = \frac{\alpha}{2}, \quad \text{and} \quad P(X \geq x_h) = \frac{\alpha}{2}.$$

Equivalently,

$$F_X(x_l) = \frac{\alpha}{2}, \quad \text{and} \quad F_X(x_h) = 1 - \frac{\alpha}{2}.$$

We can rewrite these equations by using the inverse function F_X^{-1} as

$$x_l = F_X^{-1}\left(\frac{\alpha}{2}\right), \quad \text{and} \quad x_h = F_X^{-1}\left(1 - \frac{\alpha}{2}\right).$$

We call the interval $[x_l, x_h]$ a $(1-\alpha)$ interval for X, or in other word the confidence interval.
Let us define a notation that is commonly used. For any $p \in [0,1]$, we define z_p as the real value for which

$$P(Z < z_p) = p$$

z_p can be described as corresponding to the $100(p)^{th}$ percentile of a distribution.
Visually, they look like this:

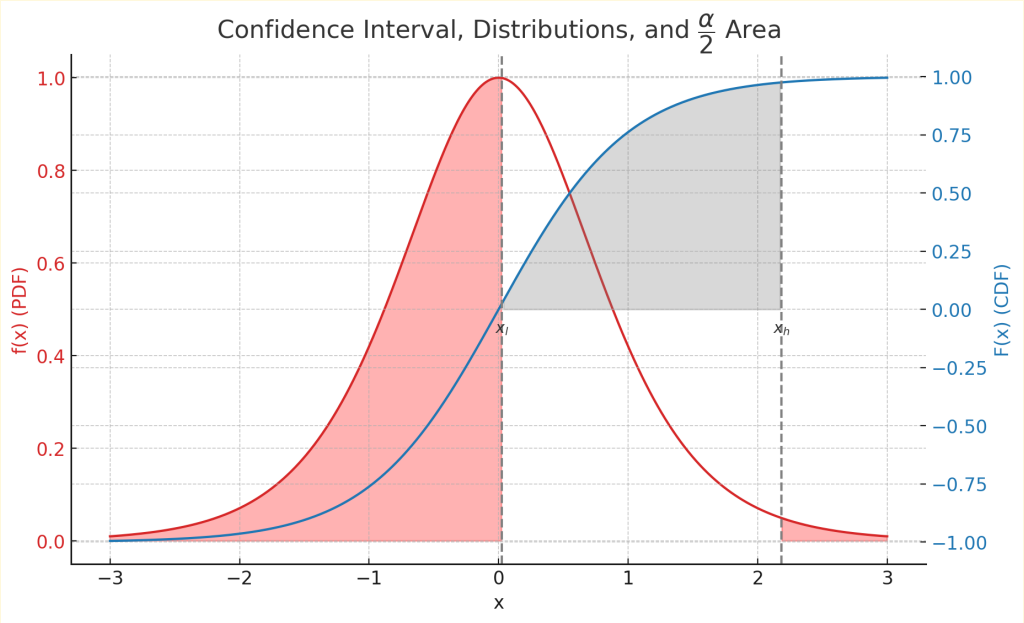


Figure 20: Visualize confidence interval

Example: Let $Z \sim N(0,1)$, find x_l and x_h such that

$$P(x_l \leq Z \leq x_h) = 0.95$$

Solution: Here, $\alpha = 0.05$ and the CDF of Z is given by the Φ function. Thus, we can choose

$$x_l = \Phi^{-1}(0.025) = -1.96, \quad \text{and} \quad x_h = \Phi^{-1}(1-0.025) = 1.96$$

Thus, for a standard normal random variable Z, we have

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

Example: This example is known as the z confidence interval. Suppose that we have a random sample $X_1, \ldots, X_n$ from the $N(\mu, \sigma^2)$ model, where $\mu \in \mathbb{R}$ is unknown and $\sigma^2 > 0$ is known.
Find $100(1-\alpha)\%$ confidence interval for μ.
Solution: To find a $100(1-\alpha)\%$ confidence interval for μ when sampling from a normal distribution $N(\mu, \sigma^2)$ where μ is unknown and σ^2 is known, we use the fact that the sample mean, $\overline{X}$, is normally distributed with $N(\mu, \frac{\sigma^2}{n})$. We have shown this fact before in the previous section.
We can standardize $\overline{X}$ to form a Z-score that follows a standard normal distribution $N(0,1)$.
The Z-score is given by:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Consider the interval $P(z_{\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}}) = 1-\alpha$.
We have:

$$P(z_{\frac{\alpha}{2}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{1-\frac{\alpha}{2}}) = 1-\alpha$$

Because standard normal distribution is symmetric, we have $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$.
Thus, after reordering the terms we have the $100(1-\alpha)\%$ confidence interval for μ is:

$$\left(\bar{X} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

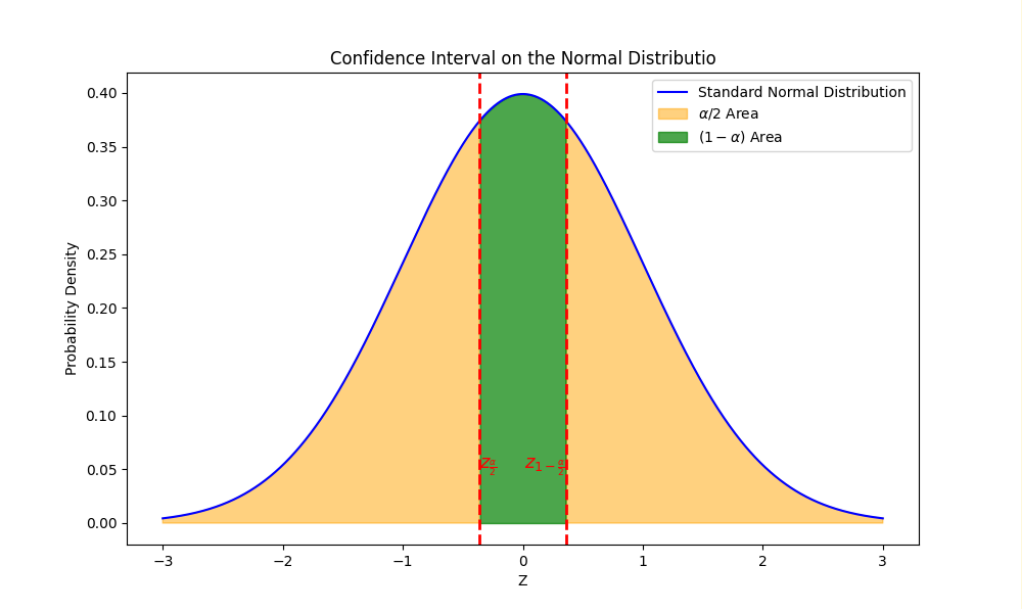


Figure 21: Visualization of the confidence interval

Example: This example is known as the t confidence interval. Consider again a random sample $X_1, \ldots, X_n$ from the $N(\mu, \sigma^2)$ model, where $\mu \in \mathbb{R}$ is unknown and $\sigma^2 > 0$ is unknown.
Find $100(1-\alpha)\%$ confidence interval for μ.
Solution: Let

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

Where $\bar{X}$ is the sample mean, and S^2 is the sample variance. We cannot use the fact that $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, because we do not know σ^2. Then

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a Student's t distribution with $n-1$ degrees of freedom. Note that the distribution of T does not depend on unobserved value $\sigma^2 > 0$ and μ. Otherwise, it is not allowed.
Recall that in the previous section, we notice that $t_{\frac{\alpha}{2}} = -t_{1-\frac{\alpha}{2}}$, it is defined similarly as the percentile.
And we have $P(t_{\frac{\alpha}{2}} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{-\frac{\alpha}{2}}) = P(-t_{1-\frac{\alpha}{2}} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{1-\frac{\alpha}{2}}) = 1-\alpha$.
And by rearrange equation, we have:

$$P\left(\bar{X} - \frac{t_{1-\frac{\alpha}{2}}S}{\sqrt{n}} < \mu < \bar{X} + \frac{t_{1-\frac{\alpha}{2}}S}{\sqrt{n}}\right) = 1-\alpha$$

Therefore, the confidence interval is:

$$\left(\bar{X} - \frac{t_{1-\frac{\alpha}{2}}S}{\sqrt{n}}, \bar{X} + \frac{t_{1-\frac{\alpha}{2}}S}{\sqrt{n}}\right)$$

Example: Now, we have more general example of normal based confidence interval. Suppose that we have a random sample $X_1, \ldots, X_n$ from a population with unknown mean μ and known variance σ^2.
Find $100(1-\alpha)\%$ confidence interval for μ.
Solution: If the sample size is sufficient large, then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. Using the Z score, we have $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$.
Similarly, we have $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$.
Then we have:

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

Rearrange the term in the equality, we have:

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

Here, the confidence interval is:

$$\left(\bar{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

Example: Now, assume we still have normal based confidence interval, but we do not know mean and variance. Suppose that we have a random sample $X_1, \ldots, X_n$ from a population with unknown mean μ and unknown variance σ^2.
Find $100(1-\alpha)\%$ confidence interval for μ.
Solution: By the example from the previous section, we proved that $Z = \frac{\bar{X}_n - \mu}{S/\sqrt{n}} \sim N(0,1)$.
As usual, we have $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$.
Then we have:

$$P\left(z_{\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}}\right) = P\left(-z_{1-\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

Rearrange the term, we have:

$$P\left(\bar{X} - z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} < \mu < \bar{X} + z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

Thus, the confidence interval is:

$$\left(\bar{X} - z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right)$$

3.6 Hypothesis testing

Definition(Hypothesis)**:** A hypothesis is a statement about a population parameter.

Definition(Null and alternative hypothesis)**:** The two complementary hypotheses in a hypothesis testing problem are called the null hypothesis and the alternative hypothesis. They are denoted by H_0 and H_1, respectively.

Remark: The goal of hypothesis test is to decide, based on a sample from the population, which of two complementary hypothesis is true.
Remark: If θ denotes a population parameter, the general format of the null and alternative hypotheses is $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_0^c$, where Θ_0 is some subset of the parameter space and Θ_0^c is its complement.

Definition(Hypothesis testing)**:** A hypothesis testing procedure or hypothesis test is a rule that specifies:

1. For which sample values the decision is made to accept H_0 as true.
2. For which sample values H_0 is rejected and H_1 is accepted as true.

The subset of the sample space for which H_0 will be rejected is called the rejection region or critical region. The complement of the rejection region is called the acceptance region.

Remark: A hypothesis testing problem is a problem in which one of two actions is going to be taken, assertion of H_0 or H_1. H_1 can also be denoted as H_a

Definition(Test statistics)**:** It is a function of the onserved sample $W(X_1, \cdots, X_n) = W(\mathbf{X})$.

Remark: Typically, a hypothesis test is specified in terms of a test statistic. We use the test statistics to help us make decisions about H_0

Definition(P-value)**:** A P-value is the probability of seeing a test statistic as or more extreme as the observed test statistic when the null is true.

Remark: This means after calculating the test statistic from your data (the "observed test statistic"), you compare this value to the distribution of test statistics that you would expect to see if the null hypothesis were true. The P-value quantifies the probability of obtaining test results as extreme as, or more extreme than, what was observed, assuming the null hypothesis is true.

Definition(Significance level α)**:** Threshold to determine if we reject H_0 based on the P-value.

Remark: If this probability is smaller than a pre-specified significance level α (usually, $\alpha = 0.1, 0.05$, or 0.01), then we reject the null.
Remark: A small P-value indicates a strong inconsistency between the observed data and what would be expected if the null hypothesis were correct. Thus, the smaller the P-value, the less likely the null is to be true.
Remark:

- An exact hypothesis test is based on a test statistic with a known distribution.
- An approximate hypothesis test is based on a test statistic with a large sample or asymptotic distribution

Example: We will discuess about Z-test. Suppose that we have a random sample $X_1, \ldots, X_n$ from the $N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ is unknown and $\sigma^2 > 0$.
Determine a procedure to calculate the P-value to test

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_0 : \mu \neq \mu_0$$

when σ^2 is known and the test statistic is

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Solution: By definition of the P-value, we know that we are looking for values of Z that are at least as extreme as the observed Z. This means we are looking for the probability of $Z > |Z_{obs}|$. This means the P-value should be:

$$P(Z > Z_{obs}) + P(Z < -Z_{obs}) = \text{P value}$$

When the P value is smaller than the selected α, it means we have enough evidence against the null hypothesis H_0, and we reject H_0 at significance level H_0.
Note when we have a smaller α, we might accept H_0, so we must be clear about α.

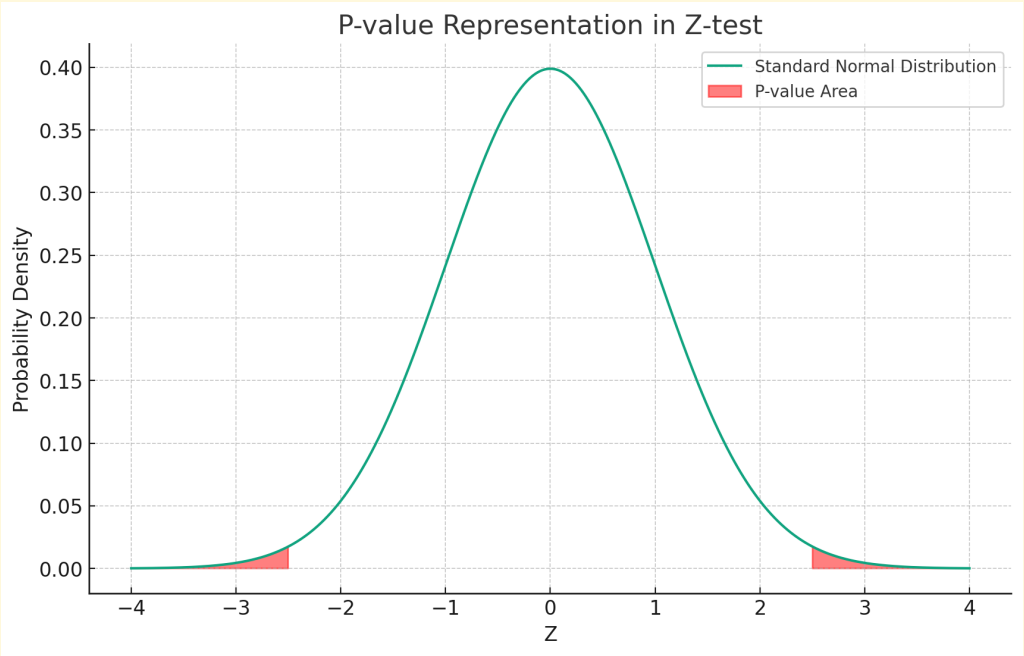


Figure 22: Graphical representation of P value

Example: Now we will explore the t-test. Suppose that we have a random sample $X_1, \ldots, X_n$ from the $N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ is unknown and $\sigma^2 > 0$ is unknown.
Determine a procedure to calculate the P-value to test

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_0 : \mu \neq \mu_0$$

when S^2 is known and the test statistic is

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{(n-1)}$$

Solution: Just as the Z-test, we have

$$P\left(T > t_{obs}\right) + P\left(T < -t_{obs}\right) = \text{P value}$$

And when the P value is smaller than the selected α, it means we have enough evidence against the null hypothesis H_0, and we reject H_0 at significance level H_0.

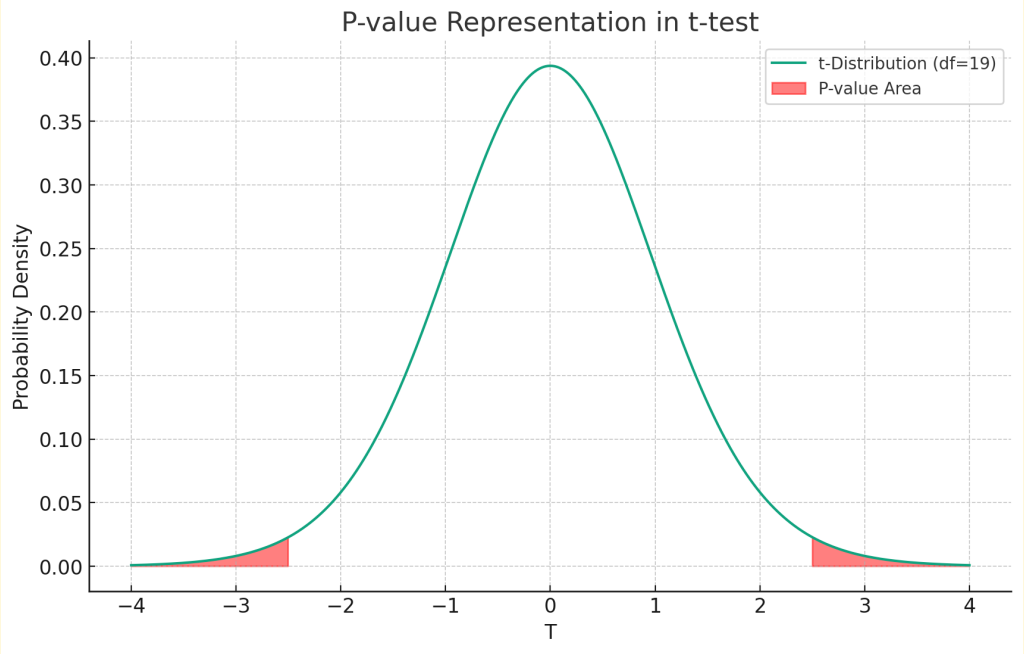


Figure 23: Visualization of p value

Example: Do you notice any similarities between the the z test and the z confidence interval? Do you think the same holds for the t test and CI?
Solution:

- Both the Z-test and the Z confidence interval rely on the standard normal distribution.

  The confidence level (e.g., 95% for Cl) complements the significance level used in hypothesis testing (e.g., $\alpha = 0.05$).

  If the null hypothesis mean (μ_0) does not fall within the calculated confidence interval, the null hypothesis would be rejected at the corresponding significance level.

  - **If μ_0 is outside the confidence interval:** This means that the range of values that you are 95% confident contains the true population mean does not include μ_0. Therefore, the sample data provide sufficient evidence to suggest that μ is significantly different from μ_0, leading to rejection of the null hypothesis at the 5% significance level.
  - **If μ_0 is inside the confidence interval:** This means that the range of values that you are 95% confident contains the true population mean includes μ_0. In this case, there's not enough evidence to reject the null hypothesis at the 5% significance level, as the hypothesized mean is consistent with the sample data.

- Both the T-test and the T confidence interval use the t-distribution.

  The relationship between confidence level and significance level observed in the Z-test and Z confidence interval also applies here.

  Similarly, if the null hypothesis mean (μ_0) is outside the T confidence interval, the null hypothesis would be rejected at the corresponding significance level.

3.7 Large sample behavior of MLE

Consistency of MLE

Theorem: Under appropriate conditions (Smoothness of the function) on the probability mass/density function $f(x;\theta)$, the maximum likelihood estimator of θ, from a random sample is consistent.

Definition(Fisher information): Let $x_1, \ldots, x_n$ be iid realizations from probability mass function $p(x;\theta)$ (if X is discrete), or from density function $f(x;\theta)$ (if X is continuous), where θ is a parameter.
The Fisher Information of the parameter θ is defined to be:

$$I(\theta) = \mathbb{E}\left[\left(\frac{d}{d\theta}\log f(\mathbf{x};\theta)\right)^2\right]$$

Under appropriate condition (Smoothness of a function):

$$I(\theta) = -\mathbb{E}\left[\frac{d^2}{d\theta^2}\log f(\mathbf{x};\theta)\right]$$

Example: Suppose random variable X has a Binomial distribution with parameters (n, p). Determine the Fisher information $I(p)$ in X.
Solution: Recall that the pmf of X is

$$p(x;p) = \begin{pmatrix} n \\ x \end{pmatrix} p^x(1-p)^{n-x}, \quad x = 0, 1, \ldots, n$$

The log of the pmf then becomes

$$\log p(x;p) = \log \begin{pmatrix} n \\ x \end{pmatrix} + x\log(p) + (n-x)\log(1-p)$$

It is easy to see that

$$\frac{d}{dp}\log p(x;p) = \frac{x}{p} - \frac{n-x}{1-p}, \quad \frac{d^2}{dp^2}\log p(x;p) = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2}$$

Hence,

$$I(p) = -E\left[\frac{d^2}{d\theta^2}\log p(X;\theta)\right] = \frac{E(X)}{p^2} + \frac{n-E(X)}{(1-p)^2} = \frac{n}{p(1-p)}$$

Example: Suppose that $X_1, \ldots, X_n$ be random sample from a Poisson distribution with parameter λ. Determine the Fisher information $I(\lambda)$.
Solution: The Probability Mass Function (PMF) of a Poisson distribution for an observation X is given by:

$$p(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots$$

The natural logarithm of the PMF, known as the log-likelihood function for an observation x, is:

$$\log p(x;\lambda) = x\log(\lambda) - \lambda - \log(x!)$$

To find the rate at which the log-likelihood changes with λ, we compute the first derivative with respect to λ :

$$\frac{d}{d\lambda}\log p(x;\lambda) = \frac{x}{\lambda} - 1$$

The second derivative of the log-likelihood with respect to λ measures the curvature of the loglikelihood function and is crucial for calculating the Fisher information. It is given by:

$$\frac{d^2}{d\lambda^2}\log p(x;\lambda) = -\frac{x}{\lambda^2}$$

Fisher information is the negative expected value of the second derivative of the log-likelihood function. Therefore, for a single observation X, the Fisher information is:

$$I_X(\lambda) = -E\left[\frac{d^2}{d\lambda^2}\log p(X;\lambda)\right] = -E\left[-\frac{X}{\lambda^2}\right] = \frac{E[X]}{\lambda^2}$$

Since $E[X] = \lambda$ for a Poisson distribution, we find that for a single observation:

$$I_X(\lambda) = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}$$

Given a random sample $X_1, \ldots, X_n$ from a Poisson distribution, the Fisher information is additive for independent observations. Thus, the total Fisher information $I(\lambda)$ for the sample is:

$$I(\lambda) = nI_X(\lambda) = \frac{n}{\lambda}$$

Approximation of sampling distribution of MLE

Theorem: Under some conditions (Smoothness) on $f(X;\theta)$, $\sqrt{nl(\theta)}\left(\hat{\theta}_{ML} - \theta\right)$ converges in distribution to a standard normal distribution.

Remark: For a iid sample, the maximum likelihood estimator approximately has a normal distribution with mean θ and variance $\frac{1}{nl(\theta)}$

Confidence interval using the MLE estimator

Directly by the theorem, for the target parameter θ with $\hat{\theta}$ being the MLE, when sample size is large:

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}, \quad \sigma_{\hat{\theta}} = \frac{1}{\sqrt{nI(\hat{\theta})}}$$

approximately has a standard normal distribution by the approximation theorem. We have:

$$P\left(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}\right) = 1 - \alpha$$

Substituting for Z in the probability statement and solving it, we obtain

$$P\left(-z_{\alpha/2} \le \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \le z_{\alpha/2}\right) = P\left(\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \le \theta \le \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}\right)$$

Thus, an approximate $100(1-\alpha)\%$ confidence interval for θ is:

$$\left(\hat{\theta}_{ML} - z_{\alpha/2}\sigma_{\hat{\theta}}, \hat{\theta}_{ML} + z_{\alpha/2}\sigma_{\hat{\theta}}\right)$$

Example: Suppose that $X_1, \ldots, X_n$ be random sample from a Poisson distribution with parameter λ. Find an approximate $100(1-\alpha)\%$ confidence interval for λ.
Solution: For a Poisson distribution, the MLE of λ, denoted as $\hat{\lambda}$, is the sample mean:

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

From the previous discussion, we determined that the Fisher information $I(\lambda)$ for a single observation from a Poisson distribution is $1/\lambda$. For a sample of size n, the Fisher information is n/λ. The variance of the MLE, $\sigma_{\hat{\lambda}}$, is given by the inverse of the square root of the Fisher information evaluated at $\hat{\lambda}$:

$$\sigma_{\hat{\lambda}} = \frac{1}{\sqrt{nI(\hat{\lambda})}} = \frac{1}{\sqrt{n/\hat{\lambda}}}$$

Using the normal approximation, we have that $Z = (\hat{\lambda} - \lambda)/\sigma_{\hat{\lambda}}$ follows a standard normal distribution for large n. Thus, the $100(1-\alpha)\%$ confidence interval for λ can be constructed using the critical value $z_{\alpha/2}$ from the standard normal distribution, where $P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) = 1 - \alpha$.
The confidence interval for λ is given by:

$$\left(\hat{\lambda} - z_{\alpha/2}\sigma_{\hat{\lambda}}, \hat{\lambda} + z_{\alpha/2}\sigma_{\hat{\lambda}}\right)$$

Substituting the expression for $\sigma_{\hat{\lambda}}$, the confidence interval becomes:

$$\left(\hat{\lambda} - z_{\alpha/2}\frac{1}{\sqrt{n/\hat{\lambda}}}, \hat{\lambda} + z_{\alpha/2}\frac{1}{\sqrt{n/\hat{\lambda}}}\right)$$

Let's simplify the expression for the confidence interval:

$$\left(\hat{\lambda} - z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{n}}, \hat{\lambda} + z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{n}}\right)$$

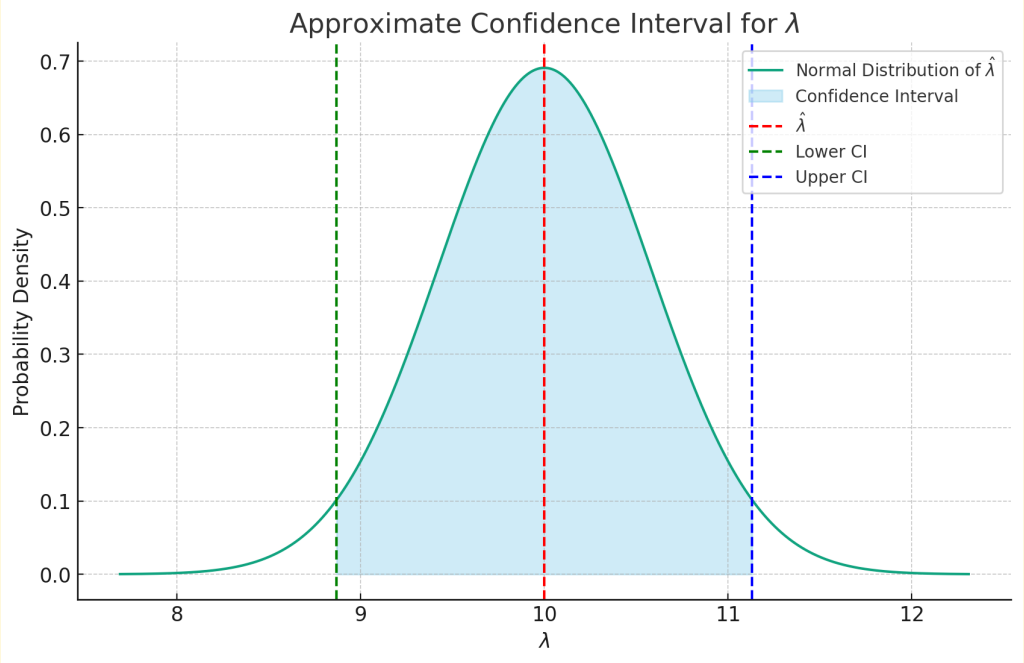


Figure 24: Visualization of the confidence interval

4 Bayesian inference

There is an unknown quantity that we would like to estimate. We get some data. From the data, we estimate the desired quantity. And in the previous chapters, we have discussed about frequentist inference to this type of the problem, which means the unknown quantity θ is assumed to be a fixed (not random) quantity that is to be estimated by the observed data. In this chapter, we will discuess a different framework of inference, it is called Bayesian inference. In the Bayesian inference, we treat the unknown quantity, θ, as a random variable. More specifically, we assume that we have some initial guess about the distribution of θ. This distribution is called the prior distribution. After observing some data, we update the distribution of θ (based on the observed data). This step is usually done using Bayes' theorem. That is why this approach is called the Bayesian inference. The details of this approach will be introduced further in the chapter.

Bayes' theorem

Theorem: Let A and B be two events. Bayes theorem states that:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

where

- $P(A \mid B)$ is the conditional probability of A given B.
- $P(B \mid A)$ is the conditional probability of B given A.
- $P(A)$ and $P(B)$ are the probabilities of A and B

4.1 The Prior and Posterior Distributions

Definition(Bayesian model): The Bayesian model for inference contains the statistical model $\{f(s \mid \theta) : \theta \in \Omega\}$ for the data $s \in S$ and adds to this the prior probability measure Π for θ.

1. **Statistical Model $\{f(s \mid \theta) : \theta \in \Omega\}$:** This part of the Bayesian model represents the set of probability distributions or likelihood functions $f(s \mid \theta)$ that describe how the data s is generated or comes about, given a set of parameters θ. The parameter space Ω includes all possible values that θ can take. Essentially, this statistical model captures the relationship between the observed data S (the sample space) and the parameters θ, allowing us to understand how likely different observations of data are under various parameter values.

2. **Prior Probability Measure Π for θ :** The Bayesian approach adds another layer to the statistical model by incorporating Π, the prior probability measure, which represents our prior knowledge or beliefs about the possible values of the parameters θ before observing any data. This prior distribution reflects our subjective assessments or objective findings from previous studies about the parameters. The prior can be specific and informative, based on substantial previous knowledge, or vague and non-informative, to minimally influence the analysis when little is known a priori.

Example: We have a motivating example on prior distribution. Suppose θ is the probability of getting a head on the toss of the coin. What is a reasonable range for θ and what type of prior might you use?
Solution: For the probability, θ, of getting a head on the toss of a coin, the range must be between 0 and 1 , inclusive. This is because $\theta = 0$ represents a situation where it is impossible to get a head (i.e., the coin always lands on tails), and $\theta = 1$ represents a situation where it is certain to get a head (i.e., the coin always lands on heads). Any value of θ between 0 and 1 represents the probability of getting a head on any given toss, with values closer to 0 indicating a lower probability and values closer to 1 indicating a higher probability.
For a probability like θ, a common choice of prior is the Beta distribution, denoted as $\mathrm{Beta}(\alpha, \beta)$, where $\alpha > 0$ and $\beta > 0$ are shape parameters.

- If $\alpha = \beta = 1$, the Beta distribution simplifies to the uniform distribution over [0, 1], representing a state of complete prior ignorance about the bias of the coin.

- If $\alpha > 1$ and $\beta = 1$, the distribution is skewed towards higher values of θ, indicating a prior belief that heads are more likely than tails.
- Conversely, if $\alpha = 1$ and $\beta > 1$, the distribution is skewed towards lower values of θ, indicating a prior belief that tails are more likely.
- If $\alpha = \beta > 1$, the distribution is symmetric around 0.5 but with a concentration indicating a belief that outcomes are likely to be fair but with some uncertainty.

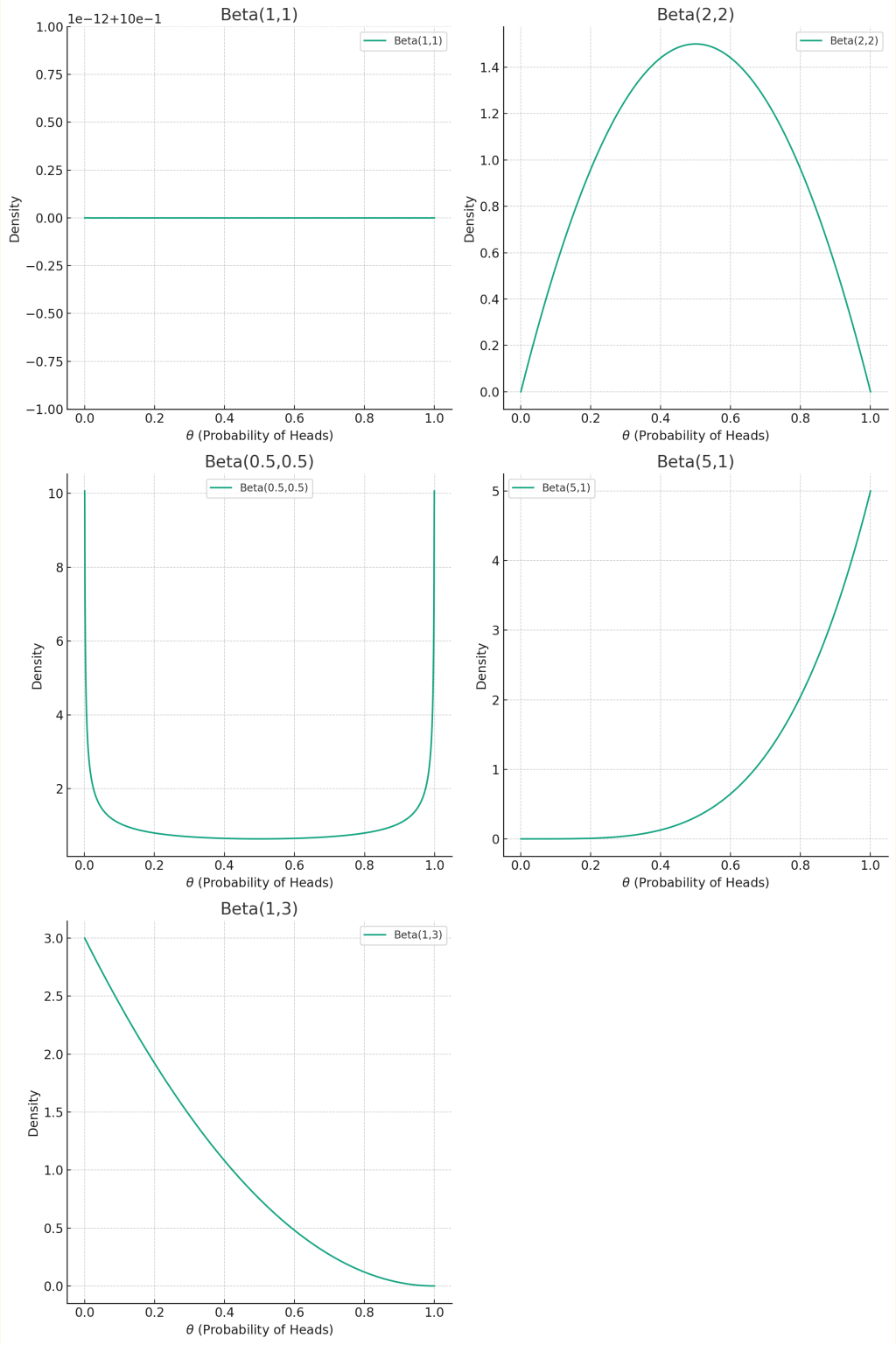


Figure 25: Different distributions leading to different theta

- $\mathrm{Beta}(1, 1)$ represents a uniform belief across all probabilities, indicating no prior bias.

- $\text{Beta}(2,2)$ shows a preference for outcomes near the middle, suggesting a belief in a fair coin but with some uncertainty.
- $\text{Beta}(0.5,0.5)$ indicates more uncertainty, with higher probabilities given to extreme outcomes (very biased towards heads or tails).
- $\text{Beta}(5,1)$ and $\text{Beta}(1,3)$ show beliefs skewed towards heads and tails, respectively, indicating a prior belief in a bias towards one side of the coin.

Definition(Prior predictive distribution): Let Ω be the parameter space. When the prior distribution is absolutely continuous, the marginal distribution for the data s is given by

$$m(s) = \int_{\Omega} \pi(\theta) f(s \mid \theta) d\theta$$

and is referred to as the prior predictive distribution of the data.

Definition(Posterior distribution): The posterior distribution of θ is the conditional distribution of θ given s. The posterior density or mass function (whichever is relevant) is given by

$$\pi(\theta \mid s) = \frac{\pi(\theta) f(s \mid \theta)}{m(s)}$$

Remark:

- The posterior density of $\theta, \pi(\theta \mid s)$, is proportional to the product of the prior density, $\pi(\theta)$, and the likelihood function, $f(s \mid \theta)$. This proportionality is expressed as:

$$\pi(\theta \mid s) \propto \pi(\theta) f(s \mid \theta)$$

- To turn the proportionality into an equality, we need a normalization factor, which ensures that the posterior distribution is a proper probability distribution that integrates to 1. This factor is $m(s)$, the marginal likelihood or evidence, also known as the prior predictive distribution. It's calculated by integrating the product of the prior and the likelihood over all possible values of θ :

$$m(s) = \int \pi(\theta) f(s \mid \theta) d\theta$$

  Essentially, $m(s)$ sums (or integrates) the weighted likelihood of the observed data across all possible values of θ, with the weights given by the prior distribution. This makes $m(s)$ a measure of how probable the observed data is under the model and prior.

- To convert the proportionality into a proper density function for θ, you divide the product of the prior and likelihood by $m(s)$:

$$\pi(\theta \mid s) = \frac{\pi(\theta) f(s \mid \theta)}{m(s)}$$

- In many practical Bayesian analyses, especially those involving computational methods like Monte Carlo simulations, we can work directly with the unnormalized posterior (i.e., $\pi(\theta) f(s \mid \theta)$) for tasks like sampling or finding modes (maximum a posteriori estimates). This is because the normalization factor, $m(s)$, does not affect the shape of the posterior distribution with respect to θ; it's merely a scaling factor to ensure the distribution integrates to 1. Thus, for many computational purposes, knowing the exact value of $m(s)$ is unnecessary, which simplifies analysis and computation.

Definition(Normalization of Beta function): The Beta function, $B(\alpha, \beta)$, is defined as an integral that integrates the function $x^{\alpha-1}(1-x)^{\beta-1}$ over the interval $[0,1]$:

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx$$

Example: Suppose we observe a sample $x_1, \ldots, x_n$ from the $\text{Bernoulli}(p)$ distribution with $p \in [0, 1]$ is unknown. For the prior, we take $\pi(\alpha, \beta)$ to be equal to a $\text{Beta}(\alpha, \beta)$ density.
Find the posterior distribution of p.
Solution: The prior distribution for p is given as a Beta distribution $\pi(p) = \text{Beta}(\alpha, \beta)$. The PDF of the Beta distribution is:

$$\pi(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)}$$

where $B(\alpha, \beta)$ is the beta function.
Given $x_1, \ldots, x_n$ are samples from a $\text{Bernoulli}(p)$ distribution, the likelihood function for observing these samples, denoted by $f(s \mid p)$, where s represents the sample data, is: $f(s \mid p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$.
The prior predictive distribution $m(s)$ is given by:

$$\begin{aligned} m(s) &= \int_0^1 \pi(p) f(s \mid p) dp \\ &= \int_0^1 \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)} \cdot p^{\sum x_i}(1-p)^{n-\sum x_i} dp \\ &= \frac{1}{B(\alpha, \beta)} \int_0^1 p^{\alpha+\sum x_i-1}(1-p)^{\beta+n-\sum x_i-1} dp \end{aligned}$$

The integral above is the normalization factor of a Beta distribution with parameters $\alpha' = \alpha + \sum x_i$ and $\beta' = \beta + n - \sum x_i$. The integral itself is equivalent to the beta function $B(\alpha', \beta')$, where $\alpha' = \alpha + \sum x_i$ and $\beta' = \beta + n - \sum x_i$. Therefore,

$$m(s) = \frac{B(\alpha + \sum x_i, \beta + n - \sum x_i)}{B(\alpha, \beta)}$$

Using the definition of the posterior distribution,

$$\pi(p \mid s) = \frac{\pi(p) f(s \mid p)}{m(s)}$$

Substituting the expressions for $\pi(p), f(s \mid p)$, and $m(s)$,

$$\begin{aligned} \pi(p \mid s) &= \frac{p^{\alpha-1}(1-p)^{\beta-1} \cdot p^{\sum x_i}(1-p)^{n-\sum x_i}}{\frac{B(\alpha+\sum x_i, \beta+n-\sum x_i)}{B(\alpha,\beta)}} \\ &= \frac{B(\alpha, \beta)}{B(\alpha + \sum x_i, \beta + n - \sum x_i)} \cdot p^{\alpha+\sum x_i-1}(1-p)^{\beta+n-\sum x_i-1} \end{aligned}$$

Since the Beta function in the denominator $B(\alpha + \sum x_i, \beta + n - \sum x_i)$ is the normalization constant for the Beta distribution with parameters $\alpha + \sum x_i$ and $\beta + n - \sum x_i$, the expression simplifies to the Beta distribution:

$$p \mid s \sim \text{Beta}\left(\alpha + \sum x_i, \beta + n - \sum x_i\right)$$

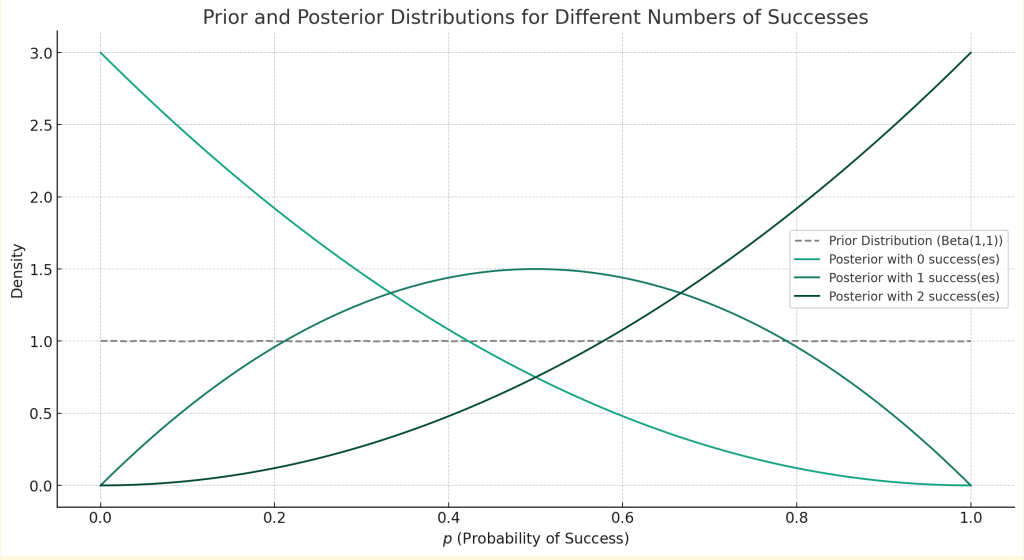


Figure 26: Different prostrior and prior distribution

Example: Suppose $x_1, \ldots, x_n$ is a random sample from $N\left(\mu, \sigma_0^2\right)$ distribution where $\mu \in \mathbb{R}$ is unknown and σ_0^2 is known.
Suppose we take the prior distribution of μ to be $N\left(\mu_0, \tau_0^2\right)$ for some specified choice of μ_0 and τ_0^2.
Find the posterior distribution of μ.
Solution: The likelihood function likelihood (data $\mid \mu$) for the data assuming it comes from a normal distribution $N\left(\mu, \sigma_0^2\right)$. The likelihood of the data given μ when $x_i \sim N\left(\mu, \sigma_0^2\right)$ is:

$$\prod_{i=1}^{n} f\left(x_i \mid \mu\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{\left(x_i-\mu\right)^2}{2\sigma_0^2}\right).$$

We simplify the product by first dealing with the constants and then with the exponents:

$$f(s \mid \mu) = \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n \prod_{i=1}^{n} \exp\left(-\frac{\left(x_i-\mu\right)^2}{2\sigma_0^2}\right).$$

The constant factor simplifies to:

$$\left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n = \frac{1}{\left(2\pi\sigma_0^2\right)^{n/2}}.$$

For the exponents, we use the property of exponents $\exp(a)\exp(b) = \exp(a+b)$:

$$\prod_{i=1}^{n} \exp\left(-\frac{\left(x_i-\mu\right)^2}{2\sigma_0^2}\right) = \exp\left(-\sum_{i=1}^{n} \frac{\left(x_i-\mu\right)^2}{2\sigma_0^2}\right)$$

Combining these, we get:

$$f(s \mid \mu) \propto \exp\left(-\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} \left(x_i-\mu\right)^2\right)$$

The sum of squared differences is:

$$\sum_{i=1}^{n} \left(x_i-\mu\right)^2.$$

Expanding each term, we get:

$$\sum_{i=1}^{n} \left(x_i^2 - 2x_i\mu + \mu^2\right).$$

Separating the sum into individual components:

$$\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2,$$

where $n\mu^2$ comes from summing μ^2 over n terms.
Thus, we have:

$$f(s \mid \mu) \propto \exp\left(n\mu^2 - 2\mu \sum_{i=1}^{n} x_i\right)$$

The prior distribution for μ, $\pi(\mu) = N\left(\mu_0, \tau_0^2\right)$:

$$\pi(\mu) \propto \exp\left(-\frac{\left(\mu-\mu_0\right)^2}{2\tau_0^2}\right) \propto \exp\left(-\frac{\left(\mu^2 - 2\mu\mu_0\right)}{2\tau_0^2}\right)$$

Now, multiply both identity we have:

$$\pi(\mu \mid s) \propto f(s)\mu)\pi(\mu) \propto \exp\left(-\frac{1}{2\sigma_0^2}\left(n\mu^2 - 2\mu\sum_{i=1}^{n} x_i\right) - \frac{1}{2\tau_0^2}\left(\mu^2 - 2\mu\mu_0\right)\right)$$
$$= \exp\left(-\frac{1}{2}\left(\left(\frac{n}{\sigma_0^2} + \frac{1}{\tau_0^2}\right)\mu^2 - 2\left(\frac{n\bar{x}}{\sigma_0^2} + \frac{\mu_0}{\tau_0^2}\right)\mu\right)\right)$$

Let:

$$\frac{1}{\sigma_*^2} = \frac{n}{\sigma_0^2} + \frac{1}{\tau_0^2}$$

and

$$\frac{\mu_*}{\sigma^2} = \frac{n\bar{x}}{\sigma_0^2} + \frac{\mu_0}{\tau_0^2}.$$

Thus:

$$\mu_* = \frac{\frac{n\bar{x}}{\sigma_0^2} + \frac{\mu_0}{\tau_0^2}}{\frac{n}{\sigma_0^2} + \frac{1}{\tau_0^2}} = \frac{n\bar{x}\tau_0^2 + \mu_0\sigma_0^2}{n\tau_0^2 + \sigma_0^2}$$

and

$$\sigma_*^2 = \left(\frac{n}{\sigma_0^2} + \frac{1}{\tau_0^2}\right)^{-1} = \frac{\sigma_0^2\tau_0^2}{n\tau_0^2 + \sigma_0^2}.$$

Hence, the posterior distribution for μ given the data is $N\left(\mu_*, \sigma_*^2\right)$, where:

$$\mu_* = \frac{n\bar{x}\tau_0^2 + \mu_0\sigma_0^2}{n\tau_0^2 + \sigma_0^2}, \quad \sigma_*^2 = \frac{\sigma_0^2\tau_0^2}{n\tau_0^2 + \sigma_0^2}.$$

In fact, if we define $k = \frac{n\tau_0^2}{n\tau_0^2+\sigma_0^2}$ then

$$\mu_* = (1-k)\mu_0 + k\bar{x} \quad \text{and} \quad \sigma_*^2 = k\frac{\sigma_0^2}{n}$$

Here k is called the Credibility Factor.

Example: Suppose we observe $x_1, \ldots, x_n$ from a Poisson (λ) distribution. Suppose we take the prior distribution of λ to be $\mathrm{Gamma}(\alpha, \beta)$ for some specified choice of α and β.
Find the posterior distribution of λ. i.e find $\pi(\lambda \mid s)$.

Solution: Given that $x_i \sim \mathrm{Poisson}(\lambda)$ for each i, the probability mass function of a single observation x_i is:

$$P\left(x_i \mid \lambda\right) = e^{-\lambda}\frac{\lambda^{x_i}}{x_i!}$$

Since the x_i are i.i.d., the joint likelihood for all observations is:

$$f(s \mid \lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}\prod_{i=1}^{n}\frac{1}{x_i!}$$

Simplifying, we focus only on the terms that involve λ :

$$f(s \mid \lambda) \propto e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}$$

The prior distribution $\lambda \sim \mathrm{Gamma}(\alpha, \beta)$ has the density:

$$f(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

where $\Gamma(\cdot)$ is the gamma function.
Multiplying the likelihood by the prior:

$$\pi(\lambda \mid s) \propto f(s \mid \lambda) \times f(\lambda) \propto e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i} \times \lambda^{\alpha-1}e^{-\beta\lambda} = e^{-(n+\beta)\lambda}\lambda\left(\sum_{i=1}^{n} x_i + \alpha - 1\right)$$

The posterior distribution of λ is also a gamma distribution, $\mathrm{Gamma}\left(\alpha + \sum_{i=1}^n x_i, \beta + n\right)$. The shape parameter α' is updated by adding the sum of the observations to the original shape parameter α, and the rate parameter β' is the original rate β plus the number of observations n :

$$\alpha' = \alpha + \sum_{i=1}^{n} x_i, \quad \beta' = \beta + n.$$

Thus, $\pi(\lambda \mid s) = \mathrm{Gamma}\left(\alpha', \beta'\right)$, where $s = \sum_{i=1}^n x_i$.

4.2 Posterior mean and posterior mode

The posterior distribution $\pi(\theta \mid s)$ contains all the relevant information about θ from the model $\{f(s \mid \theta) : \theta \in \Omega\}$, the data s and the prior π. Suppose we are interested in inferences about a real-valued characteristic of interest, $\psi(\theta)$. $\psi(\theta)$ is a real value function, which can be a identity function.

The most natural estimator of $\psi(\theta)$ is the mode of the posterior density/mass of $\psi(\theta)$. This is the point where the posterior density/mass function of ψ takes its maximum. An alternative estimate is given by the posterior mean, $E(\psi(\theta) \mid s)$ (provided it exists). When the posterior distribution of ψ is symmetric about its mode then the posterior expectation is the same as the posterior mode; otherwise, these estimates will be different. If we want the estimate to reflect where the central mass of probability lies, then in cases where the posterior is highly skewed, the mode is a better choice than the mean.

Example: Suppose we observe $x_1, \ldots, x_n$ from the $\mathrm{Bern}(p)$ distribution with $p \in [0,1]$ unknown and we usa a $\mathrm{Beta}(\alpha, \beta)$ prior on p.

In the example, we determined the posterior distribution of p to be $\mathrm{Beta}(n\bar{x} + \alpha, n(1-\bar{x}) + \beta)$. Let us suppose that the characteristic of interest is $\psi(p) = p$.

Determine the posterior mode posterior mean.

Solution: For a Beta distribution $\mathrm{Beta}(\alpha, \beta)$, the mean is given by:

$$\mathrm{Mean}(p) = \frac{\alpha}{\alpha + \beta}$$

Thus, for the posterior distribution $\mathrm{Beta}(n\bar{x} + \alpha, n(1-\bar{x}) + \beta)$, the posterior mean is:

$$\mathrm{Mean}(p \mid x) = \frac{n\bar{x} + \alpha}{n\bar{x} + \alpha + n(1-\bar{x}) + \beta} = \frac{n\bar{x} + \alpha}{n + \alpha + \beta}$$

For a Beta distribution $\mathrm{Beta}(\alpha, \beta)$, the mode (when both parameters are greater than 1) is given by:

$$\mathrm{Mode}(p) = \frac{\alpha - 1}{\alpha + \beta - 2}.$$

For the posterior Beta distribution $\mathrm{Beta}(n\bar{x} + \alpha, n(1-\bar{x}) + \beta)$, assuming $n\bar{x} + \alpha > 1$ and $n(1-\bar{x}) + \beta > 1$, the mode is:

$$\mathrm{Mode}(p \mid x) = \frac{(n\bar{x} + \alpha) - 1}{(n\bar{x} + \alpha) + (n(1-\bar{x}) + \beta) - 2} = \frac{n\bar{x} + \alpha - 1}{n + \alpha + \beta - 2}$$

Example: Suppose we observe $x_1, \ldots, x_n$ from a $N(\mu, \sigma_0^2)$ distribution, where $\mu \in \mathbb{R}$ is unknown and σ_0^2 is known, and we take the prior distribution on μ to be $N(\mu_\sigma, \tau_0^2)$.

Suppose, that the characteristic of interest is $\psi(\mu) = \mu$.

Find the posterior mode and mean.

Solution: For a normal distribution, the mean is the same as the expectation, so the posterior mean of μ is μ_n :

$$\mathrm{Mean}(\mu \mid x) = \mu_n = \frac{n\bar{x}\tau_0^2 + \mu_0\sigma_0^2}{n\tau_0^2 + \sigma_0^2}$$

For a normal distribution, the mode is also the peak of the distribution, which is the mean for a symmetric normal distribution. Thus, the posterior mode of μ is also μ_n :

$$\mathrm{Mode}(\mu \mid x) = \mu_n = \frac{n\bar{x}\tau_0^2 + \mu_0\sigma_0^2}{n\tau_0^2 + \sigma_0^2}$$

4.3 Credible interval

Definition(Credible interval)**:** A credible interval, for a real-valued parameter $\psi(\theta)$, is an interval $C(s) = [I(s), u(s)]$ that will contain the true value of the parameter with some amount of certainty. A γ-credible interval is an interval $C(s)$ that satisfies

$$\Pi(\psi(\theta) \in C(s) \mid s) \geq \gamma$$

Where Π means the probability.

Definition(Highest posterior density)**:** A γ-HPD interval for a real-valued parameter θ is an interval or set of intervals $C(s) = \theta \in \Theta : \pi(\theta \mid s) \geq k$ that satisfies two conditions:

- The interval $C(s)$ covers at least a γ proportion of the probability mass of the posterior distribution, i.e.

$$\Pi(\theta \in C(s) \mid s) \geq \gamma$$

  where Π denotes the posterior probability.

- The threshold k is the largest number such that the set $C(s)$ defined by the condition $\pi(\theta \mid s) \geq k$ contains γ of the posterior distribution. This condition ensures that $C(s)$ includes the densest regions of the posterior distribution, thereby minimizing the measure (or length) of $C(s)$ while still containing the specified probability mass:

$$\int_{\theta:\pi(\theta|s)\geq k} \pi(\theta \mid s) d\theta = \gamma.$$

Example: Suppose we observe $x_1, \ldots, x_n$ from a $N(\mu, \sigma_0^2)$ distribution, where $\mu \in \mathbb{R}$ is unknown and σ_0^2 is known, and we take the prior distribution on μ to be $N(\mu_0, \tau_0^2)$.
Suppose, that the characteristic of interest is $\psi(\mu) = \mu$.
Find the highest posterior density interval for μ.
Solution: As previously derived, given the likelihood from a normal distribution and a normal prior, the posterior distribution of μ is also normal. The posterior parameters μ_n (mean) and σ_n^2 (variance) are given by:

$$\mu_n = \frac{n\bar{x}\tau_0^2 + \mu_0\sigma_0^2}{n\tau_0^2 + \sigma_0^2}$$
$$\sigma_n^2 = \frac{\sigma_0^2\tau_0^2}{n\tau_0^2 + \sigma_0^2}$$

where $\bar{x}$ is the sample mean of the observed data, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
The posterior distribution of μ can thus be written as:

$$\mu \mid x_1, \ldots, x_n \sim N\left(\mu_n, \sigma_n^2\right)$$

Determine the z-value from the standard normal distribution that corresponds to the desired cumulative probability $\frac{\gamma+1}{2}$ (to account for both tails). For instance, for a 95% HPD interval, we look for the z-value that gives us 0.975 in the cumulative standard normal distribution, which is approximately 1.96 . Multiply this z-value by the square root of the posterior variance to get the radius of the interval from the mean:

$$\text{Radius } = z \cdot \sqrt{\sigma_n^2}$$

Where z is the z-value corresponding to $\frac{\gamma+1}{2}$ in the standard normal distribution. The HPD interval is then:

$$\left[\mu_n - z \cdot \sqrt{\sigma_n^2}, \mu_n + z \cdot \sqrt{\sigma_n^2}\right]$$

5 Optamality

5.1 Optimal estimator

In most problems, there are a variety of possible estimators for a parameter. We cannot evaluate how "good" a point estimation procedure is on the basis of the value of a single estimate. We need to consider the sampling distribution of the estimator.

Again consider $X_1, \ldots, X_n \sim^{\text{iid}} f(x;\theta)$. We want to estimate some real-valued characteristic $\psi(\theta)$ for the statistical model $\{f(x;\theta) : \theta \in \Theta\}$. $T = g(X_1, \ldots, X_n)$ is an estimator of $\psi(\theta)$ Our goal is to find an optimal estimator of $\psi(\theta)$.

We'll consider estimators that are optimal with respect to MSE. We will do this by augmenting an estimator T with a sufficient statistic U to lower its MSE.

An equivalent definition of sucient statistic

Theorem: A statistic U is sufficient for a model if and only if the conditional distribution of the data s given $U = u$ is the same for every $\theta \in \Theta$.

Example: Let $X_1, X_2, \ldots, X_n$ be a random sample form the Bernoulli distribution.
Suppose that we are given a value of $U = \sum_{i=1}^{n} X_i$, which is the number of successes in n trials.
Show that U is a sufficient statistic for p.
Solution: Given a random sample $X_1, X_2, \ldots, X_n$ from a Bernoulli distribution with parameter p, the joint probability distribution, as expressed previously, is:

$$P(X_1 = x_1, \ldots, X_n = x_n) = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i} = p^U(1-p)^{n-U}$$

where $U = \sum_{i=1}^{n} x_i$ is the number of successes. The random variable U is the sum of n independent Bernoulli random variables, each with parameter p. Hence, U follows a binomial distribution with parameters n and p, expressed as:

$$P(U = u) = \begin{pmatrix} n \\ u \end{pmatrix} p^u(1-p)^{n-u}$$

To find $P(X_1 = x_1, \ldots, X_n = x_n \mid U = u)$, we use the definition of conditional probability:

$$P(X_1 = x_1, \ldots, X_n = x_n \mid U = u) = \frac{P(X_1 = x_1, \ldots, X_n = x_n, U = u)}{P(U = u)}$$

Given $U = u$, it necessarily holds that $\sum_{i=1}^{n} x_i = u$. Thus, the numerator is simply the probability of any sequence where $\sum_{i=1}^{n} x_i = u$, which is $p^U(1-p)^{n-U}$.
The key insight here is that the numerator $p^U(1-p)^{n-U}$ equals the marginal probability of $U = u$ since U is exactly the sum $\sum_{i=1}^{n} x_i = u$. Therefore, all $(x_1, \ldots, x_n)$ configurations leading to $U = u$ will have the same joint probability distribution value.
Now, using the expression for $P(U = u)$ in the denominator:

$$P(X_1 = x_1, \ldots, X_n = x_n \mid U = u) = \frac{p^u(1-p)^{n-u}}{\begin{pmatrix} n \\ u \end{pmatrix} p^u(1-p)^{n-u}} = \frac{1}{\begin{pmatrix} n \\ u \end{pmatrix}}$$

The result $\frac{1}{\begin{pmatrix} n \\ u \end{pmatrix}}$ indicates that the probability of observing any particular sequence of $X_1, \ldots, X_n$ leading to $U = u$ is equally likely, and crucially, it does not depend on the parameter p. This uniform distribution over all sequences that sum to u shows that knowing $U = u$ gives us all the information about p that could be gleaned from the data, and no additional information about p can be obtained by knowing the specific sequence of $x_1, \ldots, x_n$. Therefore, U is a sufficient statistic for p.

Rao–Blackwell theorem

Theorem: Suppose that T is an estimator of $\psi(\theta), U$ is a sufficient statistic for θ, and $E\left(T^2\right)$ is finite for every $\theta \in \Theta$. Let $\tilde{T} = E(T \mid U)$. Then

$$\mathrm{MSE}(\tilde{T}) \leq \mathrm{MSE}(T)$$

for every $\theta \in \Theta$.
The theorem states that if you have an estimator T of $\psi(\theta)$ and a sufficient statistic U for θ, then you can construct a new estimator $\tilde{T} = E(T \mid U)$, which is the conditional expectation of T given U. This new estimator $\tilde{T}$ will never have a higher MSE than T for any value of θ. Thus, $\tilde{T}$ is a more optimal estimator than T.

Theorem (For unbiased estimator)**:** Suppose that T has finite second moment and is unbiased for $\psi(\theta)$. Let U be a sufficient statistic.
Let $\tilde{T} = E(T \mid U)$, then

- $E(\tilde{T}) = \psi(\theta)$
- $\mathrm{Var}(\tilde{T}) \leq \mathrm{Var}(T)$

This theorem is a more specific application of the Rao-Blackwell theorem that explicitly states and uses the unbiasedness of T. The general Rao-Blackwell theorem focuses more broadly on the mean squared error (MSE), which includes both variance and bias components. In the specific case where T is unbiased, the MSE simplifies to the variance.

Proof. ■

Proof. ■

Example: Let $X_1, X_2, \ldots, X_n$ be a random sample from Bernoulli distribution with parameter p. Show that

$$T = \begin{cases} 1 & \text{if } X_1 = 1 \text{ and } X_2 = 0 \\ 0 & \text{Otherwise} \end{cases}$$

is unbiased for estimating $p(1-p)$ and use a sufficient statistic, $\sum_{i=1}^{n} X_i$, to find an estimator with lower variance than T.
Solution: The estimator T is defined as:

$$T = \begin{cases} 1 & \text{if } X_1 = 1 \text{ and } X_2 = 0 \\ 0 & \text{otherwise} \end{cases}$$

We need to find the expected value of T and show that it equals $p(1-p)$. Since X_1 and X_2 are independent Bernoulli random variables with parameter p, we have:

$$P\left(X_1 = 1 \text{ and } X_2 = 0\right) = P\left(X_1 = 1\right) P\left(X_2 = 0\right) = p(1-p)$$

Since T is 1 only when $X_1 = 1$ and $X_2 = 0$, and 0 otherwise, the expected value of T is simply:

$$E(T) = 1 \cdot P\left(X_1 = 1 \text{ and } X_2 = 0\right) + 0 \cdot P(\text{ otherwise }) = p(1-p)$$

Hence, T is an unbiased estimator for $p(1-p)$.
Given that $\sum_{i=1}^{n} X_i$ is a sufficient statistic for p in a Bernoulli distribution, let's denote $U = \sum_{i=1}^{n} X_i$.
To reduce the variance of T, we use the Rao-Blackwell theorem and find $\tilde{T} = E(T \mid U)$. Let's compute $\tilde{T}$:

$$E(T \mid U = u) = P\left(X_1 = 1 \text{ and } X_2 = 0 \mid U = u\right)$$

Given $U = u$, the total number of 1's in the sample n is u. The probability that $X_1 = 1$ and $X_2 = 0$ specifically given $U = u$ depends on the number of ways to arrange the remaining $u-1$ ones among $n-2$ remaining trials.

This gives us:

$$P\left(X_1 = 1 \text{ and } X_2 = 0 \mid U = u\right) = \frac{\binom{n-2}{u-1}}{\binom{n}{u}}$$

This simplifies to:

$$\frac{(n-2)!/(u-1)!(n-u-1)!}{n!/(u!(n-u)!)} = \frac{u(n-u)}{n(n-1)}$$

Therefore, $\tilde{T} = \frac{U(n-U)}{n(n-1)}$ as an estimator based on the sufficient statistic U.
The variance of $\tilde{T}$ can be computed, and it will generally be less than the variance of T due to the Rao-Blackwell theorem. Specifically, $\tilde{T}$ leverages the entire sample, not just X_1 and X_2, thereby utilizing more information from the data, leading to a reduction in variance.
To sum up, $\tilde{T} = \frac{U(n-U)}{n(n-1)}$ is an estimator of $p(1-p)$ based on the sufficient statistic U and has lower variance than the initial estimator T.

Definition(Minimum-variance unbiased estimation)**:** An unbiased estimator of $\psi(\theta)$ with smallest variance for each $\theta \in \Theta$ is called a uniformly minimum variance unbiased (UMVU) estimator.

The Cramer-Rao inequality

Theorem: Consider the model $\{f(x;\theta) : \theta \in \Theta\}$. Suppose that T is an unbiased estimator of θ based on $X_1, \ldots, X_n$.
Then,

$$\text{Var}(T) \geq \frac{1}{nI(\theta)}$$

where $I(\theta)$ is the Fisher information.

MLEs acheive this bound asymptotically and are called asymptotically efficient estimators.
The Rao-Blackwell theorem provides a method to lower the variance of an unbiased estimator.
The Cramer-Rao inequality provides a lower bound on the variance of an unbiased estimator (or variance of the UMVU estimator).

Example: Suppose that $X_1, \ldots, X_n \sim^{\text{iid}}$ Poisson (λ). Is $\bar{X}$ an asymptotically efficient estimator of λ ? Why or why not?
Solution: For a Poisson distribution with mean and variance λ, the probability mass function of a single observation X_i is:

$$f\left(x_i;\lambda\right) = e^{-\lambda}\frac{\lambda^{x_i}}{x_i!}, \quad x_i = 0, 1, 2, \ldots$$

The Fisher information $I(\lambda)$ for a single observation from this distribution is given by:

$$I(\lambda) = E\left[\left(\frac{\partial}{\partial\lambda}\log f\left(X_i;\lambda\right)\right)^2\right]$$

Calculating the log of the PMF:

$$\log f\left(x_i;\lambda\right) = -\lambda + x_i\log\lambda - \log\left(x_i!\right)$$

Taking the derivative with respect to λ:

$$\frac{\partial}{\partial\lambda}\log f\left(x_i;\lambda\right) = -1 + \frac{x_i}{\lambda}$$

So, the Fisher information becomes:

$$I(\lambda) = E\left[\left(-1 + \frac{X_i}{\lambda}\right)^2\right] = E\left[\left(\frac{X_i - \lambda}{\lambda}\right)^2\right] = \frac{1}{\lambda^2}E\left[(X_i - \lambda)^2\right] = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}$$

Given n iid observations, the Fisher information for the sample mean $\bar{X}$ is:

$$I_n(\lambda) = nI(\lambda) = \frac{n}{\lambda}$$

Thus, the Cramér-Rao lower bound for the variance of any unbiased estimator T of λ based on the sample is:

$$\mathrm{Var}(T) \geq \frac{1}{nI(\lambda)} = \frac{\lambda}{n}$$

$\bar{X}$ is an unbiased estimator of λ, with variance:

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(X_i\right) = \frac{1}{n^2}\cdot n\lambda = \frac{\lambda}{n}$$

This matches the Cramér-Rao lower bound.
Since $\bar{X}$ achieves the Cramér-Rao lower bound, it is the UMVU estimator of λ. Additionally, by achieving this bound, $\bar{X}$ is asymptotically efficient. MLEs are asymptotically efficient, and since $\bar{X}$ is the MLE of λ for the Poisson distribution, it indeed achieves the efficiency asymptotically. Thus, $\bar{X}$ is an asymptotically efficient estimator of λ.

5.2 Optimal hypothesis testing

We would like to introduce a relatively general hypothesis testing procedure called the likelihood ratio test. In this course, we'll focus on simple hypothesis tests of the form

$$H_0 : \theta = \theta_0 \text{ vs. } H_A : \theta = \theta_1$$

Likelihood ratio test (LRT) of a simple hypothesis

Theorem: Let $L\left(\theta; x_1, \ldots, x_n\right)$ be the likelihood function. The LRT of

$$H_0 : \theta = \theta_0 \quad \text{vs. } H_A : \theta = \theta_1$$

rejects H_0 if the likelihood ratio

$$\lambda\left(x_1, \ldots, x_n\right) = \frac{L\left(\theta_0; x_1, \ldots, x_n\right)}{L\left(\theta_1; x_1, \ldots, x_n\right)} < k$$

for some value k called the critical value, which is determined by the significant level α.

Remark: If λ is close to 1 , it means that the data are about equally likely under both hypothesesn either hypothesis is strongly favored.
If λ is much less than 1 , it indicates that the likelihood of the data under θ_0 is much smaller than under θ_1. This suggests that the data are less compatible with H_0 compared to H_A, making θ_1 a more plausible parameter value than θ_0 given the data.
Remark: You reject H_0 in favor of H_A when the likelihood ratio is less than a critical value k. This decision rule implies that H_0 is rejected when the evidence (in terms of likelihood) supporting H_A is sufficiently stronger compared to H_0. The choice of k controls how much stronger the evidence needs to be to justify rejecting H_0.
Remark: The critical region C for the LRT is determined by:

$$C = \{(x_1, \ldots, x_n) \mid \lambda\left(x_1, \ldots, x_n\right) < k\}$$

where k is a threshold value that defines when the null hypothesis should be rejected.

Example: Suppose that $X_1, \ldots, X_n$ are a random sample from an exponential distribution with parameter $1/\theta$. We want to test

$$H_0 : \theta = 2 \text{ vs. } H_A : \theta = \theta_1$$

where $\theta_1 > 2$.
Find the likelihood ratio and form of the LRT.

Solution: The probability density function for a single observation from an exponential distribution with rate $1/\theta$ is given by:

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}, \quad x \geq 0$$

Given n independent and identically distributed observations from this distribution, the joint likelihood function for the sample $X_1, X_2, \ldots, X_n$ is:

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-x_i/\theta} = \theta^{-n} e^{-\sum_{i-1}^{n} x_i/\theta}$$

The likelihood ratio $\lambda(x_1, \ldots, x_n)$ for testing $\theta = 2$ against $\theta_1 > 2$ is:

$$\lambda(x_1, \ldots, x_n) = \frac{L(\theta_0; x_1, \ldots, x_n)}{L(\theta_1; x_1, \ldots, x_n)} = \frac{2^{-n} e^{-\sum_{i-1}^{n} x_i/2}}{\theta_1^{-n} e^{-\sum_{i-1}^{n} x_i/\theta_1}}$$

Simplifying the ratio yields:

$$\lambda(x_1, \ldots, x_n) = \left(\frac{2}{\theta_1}\right)^n e^{-\sum_{i=1}^{n} x_i(1/2 - 1/\theta_1)}$$

Given $\theta_1 > 2, 1/2 - 1/\theta_1 < 0$. Thus, the likelihood ratio depends both on the ratio $\frac{2}{\theta_1}$ and the exponent involving the sum of the x_i 's.

Definition(Type I error): A Type I error occurs when the null hypothesis (H_0) is true, but the test incorrectly rejects it. This is also known as a "false positive" error. The probability of making a Type I error is denoted by α, which is also known as the significance level of the test.

Definition(Type II error): A Type II error occurs when the null hypothesis (H_0) is false, but the test fails to reject it. This error is also referred to as a "false negative." The probability of correctly rejecting a false null hypothesis is known as the power of the test (1 minus the probability of a Type II error, denoted as β). Increasing the power of a test reduces the likelihood of a Type II error. It is affected by the sample size, significance level, and the true value of the parameter being tested.

Example: Suppose that $X_1, \ldots, X_n$ are a random sample from an exponential distribution with parameter $1/\theta$. We want to test

$$H_0 : \theta = 2 \text{ vs. } H_A : \theta = \theta_1$$

where $\theta_1 > 2$.
Find the critical value and region for the likelihood ratio test with $\alpha = 0.05$.
Solution: For a sample $X_1, \ldots, X_n$ from an exponential distribution with parameter $1/\theta$, the probability density function for each X_i is:

$$f(x_i;\theta) = \frac{1}{\theta} e^{-x_i/\theta}, \quad x_i \geq 0$$

The likelihood function for the entire sample is:

$$L(\theta; x_1, \ldots, x_n) = \left(\frac{1}{\theta}\right)^n e^{-\sum_{i=1}^{n} x_i/\theta}$$

The likelihood ratio comparing H_0 and H_A is:

$$\lambda(x_1, \ldots, x_n) = \frac{L(2; x_1, \ldots, x_n)}{L(\theta_1; x_1, \ldots, x_n)} = \left(\frac{2}{\theta_1}\right)^n e^{-\left(\frac{1}{\theta_1} - \frac{1}{2}\right)\sum_{i=1}^{n} x_i}$$

Using the Type I error formula, we know:

$$\alpha = 0.05 = P\left(\left(\frac{2}{\theta_1}\right)^n e^{-\left(\frac{1}{\theta_1} - \frac{1}{2}\right)\sum_{i=1}^{n} x_i} < k \mid H_0 \text{ is true}\right)$$

Because each $x_1, x_2, \cdots, x_n \overset{i.i.d}{\sim} \exp\left(\frac{1}{\theta}\right)$, we have $n\overline{x} \sim \text{Gamma}\left(n, \frac{1}{\theta}\right)$.
And we have:

$$\begin{aligned}
0.05 &= P\left(\left(\frac{2}{\theta_1}\right)^n e^{-\left(\frac{1}{\theta_1}-\frac{1}{2}\right)\sum_{i=1}^n x_i} < k \mid H_0 \text{ is true}\right) \\
&= P\left(e^{-\left(\frac{1}{\theta_1}-\frac{1}{2}\right)\sum_{i=1}^n x_i} < \left(\frac{2}{\theta_1}\right)^{-n} k \mid H_0 \text{ is true}\right) \\
&= P\left(\left(\frac{1}{\theta_1}-\frac{1}{2}\right)\sum_{i=1}^n x_i < \log(k) - n\log\left(\frac{2}{\theta_1}\right) \mid H_0 \text{ is true}\right)
\end{aligned}$$

Since $\theta_1 > 2, \frac{1}{\theta_1} < \frac{1}{2}$. Thus, the term $\left(\frac{1}{\theta_1} - \frac{1}{2}\right)$ is negative. And we have:

$$0.05 = P\left(\sum_{i=1}^n x_i < \frac{\log(k) - n\log\left(\frac{2}{\theta_1}\right)}{\left(\frac{1}{\theta_1}-\frac{1}{2}\right)} \mid \theta = 2\right)$$

Thus, to solve for k, we only need to consider the quantiles of $\text{Gamma}\left(n, \frac{1}{\theta}\right)$, and $\theta = 2$:

$$\frac{\log(k) - n\log\left(\frac{2}{\theta_1}\right)}{\left(\frac{1}{\theta_1}-\frac{1}{2}\right)} = q^{0.95}_{\text{Gamma}\left(n,\frac{1}{2}\right)}$$

Thus, $k = \exp\left(\text{Gamma}\left(n, \frac{1}{2}\right)\right)\left(\frac{1}{\theta_1} - \frac{1}{2} + n\log\left(\frac{2}{\theta_1}\right)\right)$

Neyman-Pearson Lemma

Lemma: Among all tests that have a fixed level α (probability of Type I error), the likelihood ratio test (LRT) that rejects H_0 when the likelihood ratio

$$\lambda(x) = \frac{L(\theta_0; x)}{L(\theta_1; x)}$$

is less than or equal to a certain constant k is the most powerful test for testing:

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_A : \theta = \theta_1$$

where θ_0 and θ_1 are specific values of the parameter θ.

Remark: The power of a test is $= P($ Reject $H_0 \mid H_A$ is true $)$.

Example: Suppose that $X_1, \ldots, X_n$ are a random sample from an exponential distribution with parameter $1/\theta$. We want to test

$$H_0 : \theta = 2 \text{ vs. } H_a : \theta = 3.$$

Is the test we developed the most powerful test at $\alpha = 0.05$ for testing against this alternative hypothesis?
Solution: According to the Neyman-Pearson Lemma, the test that rejects H_0 when $\lambda(x_1, \ldots, x_n) \leq k$ is the most powerful test for these simple hypotheses at the given α. You must find k such that:

$$P\left(\left(\frac{2}{3}\right)^n e^{-\frac{1}{6}\sum_{i-1}^n X_i} \leq k \middle| \theta = 2\right) = 0.05$$

6 Linear regression

6.1 Model formulations

Definition(Regression framework)**:** Regression is a framework for modeling a random variable (outcome) as a function of other random variables (covariates).

Definition(Regression function)**:** Let Y denote the outcome and $(X_1, \ldots, X_p)$ the covariates. The regression function

$$E\left(Y \mid X_1 = x_1, \ldots, X_p = x_p\right)$$

models the relationship between the outcome and covariates.

Definition(Linear regression model)**:** Assume that Y is a continuous random variable and X a covariate. The simple linear regression model is given by:

$$E(Y \mid X = x) = \beta_0 + \beta_1 x \quad \text{where}$$

- β_0 and β_1 are fixed, but unknown parameters (or coefficients).
- β_0 is the intercept and β_1 is the slope.

Remark: If the regression model includes a scenario where it makes sense for X to be zero (and $X = 0$ is within the range of your data), β_0 can be interpreted directly as the average outcome of Y for $X = 0$.
Remark: β_1 measures how much Y is expected to increase (or decrease, if β_1 is negative) when X increases by one unit. Positive β_1 indicates a positive association between X and Y; as X increases, Y also increases. Negative β_1 indicates a negative association; as X increases, Y decreases. Magnitude of β_1 means the larger the absolute value of β_1, the stronger the influence of X on Y.

Definition(Another form of linear regression)**:** The simple linear regression model is often written as:

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where ϵ is an independent random error with $E(\epsilon \mid X = x) = 0$ and $\mathrm{Var}(\epsilon \mid X = x) = \sigma^2$.

Remark: The expectation of Y given $X = x$ can be derived by taking the expected value of both sides of the equation:

$$E(Y \mid X = x) = E\left(\beta_0 + \beta_1 x + \epsilon \mid X = x\right)$$

Since β_0 and β_1 are constants (not random variables), and x is given, their expected values are themselves. Thus, the equation simplifies to:

$$E(Y \mid X = x) = \beta_0 + \beta_1 x + E(\epsilon \mid X = x)$$

Given that $E(\epsilon \mid X = x) = 0$ (by model assumption), the equation further simplifies:

$$E(Y \mid X = x) = \beta_0 + \beta_1 x + 0 = \beta_0 + \beta_1 x$$

Remark: The variance of Y given $X = x$ can be derived by applying the properties of variance to the model equation:

$$\mathrm{Var}(Y \mid X = x) = \mathrm{Var}\left(\beta_0 + \beta_1 x + \epsilon \mid X = x\right)$$

Again, since β_0 and β_1 are constants and x is given, their contribution to the variance is zero. Hence, the equation simplifies to:

$$\mathrm{Var}(Y \mid X = x) = \mathrm{Var}(\epsilon \mid X = x)$$

Given that $\mathrm{Var}(\epsilon \mid X = x) = \sigma^2$ (by model assumption), we then have:

$$\mathrm{Var}(Y \mid X = x) = \sigma^2$$

6.2 Parameter estimation

We do not know the values of the parameters in the population, so we use our sample to estimate them. The goal of parameter estimation is to find the values of β_0 and β_1 that "best fit" our data. We call these values estimates of the model parameters and denote them as $\hat{\beta}_0$ and $\hat{\beta}_1$. There are many possible lines that can be fit.

Suppose the data consists of $(y_1, x_1), \ldots, (y_n, x_n)$. $\star$ Given estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, the predicted (or fitted) values of the outcome are

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$\star$ The differences between the true and predicted values of the outcome are called the residuals,

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$$

Parameter estimation with least squares

Theorem: The least squares estimates are the values of β_0 and β_1 that minimize the residual sum of squares and are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \text{ and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

A Important R code used in examples

```r
# Example 1 in Sampling distribution
# Set seed for reproducibility
set.seed(123)

# Step 1: Generate a population from a normal distribution
population <- rnorm(10000, mean = 50, sd = 10)

# Step 2 and 3: Draw 1000 samples of size 100 and calculate their means
sample_means <- replicate(1000, mean(sample(population, size = 100, replace = TRUE)))

# Step 4: Plot the sampling distribution of the sample mean
hist(sample_means, breaks = 30, col = "blue", main = "Sampling Distribution of the Sample Mean", xlab = "Sample Mean")

# Adding more details to the plot for clarity
abline(v = mean(sample_means), col = "red", lwd = 2)
legend("topright", legend = paste("Mean =", round(mean(sample_means), 2)), col = "red", lwd = 2)
```

```r
# Weak law of large numbers visualizations.
set.seed(123) # For reproducibility

# Number of flips and trials
n_flips <- 1000

# Simulate coin flips: 1 for heads, 0 for tails
coin_flips <- sample(c(0, 1), size = n_flips, replace = TRUE, prob = c(0.5, 0.5))

# Calculate cumulative mean of heads
cumulative_means <- cumsum(coin_flips) / (1:n_flips)

# Plot the cumulative mean of heads against the number of flips
plot(1:n_flips, cumulative_means, type = "l", col = "blue", xlab = "Number of Flips", ylab = "Fraction of Heads",
     main = "Weak Law of Large Numbers: Fraction of Heads in Coin Flips")
abline(h = 0.5, col = "red", lwd = 2)

# Add a legend
legend("bottomright", legend = c("Cumulative Fraction of Heads", "Expected Fraction (0.5)"),
       col = c("blue", "red"), lwd = 2, bty = "n")
```

```r
# CLT simulation
# Load necessary library
library(ggplot2)

# Function to simulate the experiment and plot histograms
simulate_clt <- function(sample_sizes, num_repeats) {
  par(mfrow = c(2, 3)) # Setting up the plot area to display multiple histograms

  for (n in sample_sizes) {
    sample_means <- replicate(num_repeats, mean(runif(n)))

    # Create a histogram for the current sample size
```

```
13      hist(sample_means, breaks = 30, main = paste("Sample Size", n)
            ,
14            xlab = "Sample Mean", ylab = "Frequency", col = "
                lightblue")
15    }
16  }
17
18  # Define sample sizes and number of repetitions
19  sample_sizes <- c(1, 2, 3, 5, 10, 30)
20  num_repeats <- 1000
21
22  # Simulate and create histograms
23  simulate_clt(sample_sizes, num_repeats)
```