LECTURE NOTES FOR STAT 381

# DESCRIPTIVE STATISTICS VS. INFERENTIAL STATISTICS

**Descriptive statistics** deals with visualization and summarization of data (observations). Data may be categorical (qualitative, different categories) or numeric (numbers on which arithmetic operations make sense). For categorical data, bar graphs and pie charts are typically constructed. For numerical data, histograms, and box plots are constructed and descriptive statistics are computed such as mean, median, mode, variance, and standard deviation.

**Inferential statistics** is concerned with the estimation of population parameters based on observed data.

In this course, we focus exclusively on inferential statistics. We will study methods of parameter estimation, properties of those estimators, interval estimators, and hypotheses testing.

# REVIEW OF PROBABILITY THEORY

## Discrete Distributions

**Definition.** A discrete random variable $X$ assumes finite or countably infinite number of values. The probability distribution of $X$ is defined by the **probability mass function** (pmf) $p_X(x) = \mathbb{P}(X = x)$.

**Bernoulli distribution.** A random variable $X$ assumes value 1 with probability $p$, and 0 with probability $1-p$. These values are sometimes termed "yes/no", or "head/tail", or "success/failure". The distribution of $X$ is $Bernoulli(p)$ where $p$ is termed the **probability of a success**. We write $X \sim Ber(p)$. The pmf of $X$ is $p_X(x) = \mathbb{P}(X = x) = p^x(1-p)^{1-x}$, $x = 0, 1$. The mean of $X$ is $\mathbb{E}X = (1)(p) + (0)(1-p) = p$. The variance of $X$ is $\mathbb{V}ar(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = (1)^2(p) + (0)^2(1-p) - p^2 = p - p^2 = p(1-p)$.

**Note.** The Bernoulli distribution is named after Jacob Bernoulli(1654 - 1705), a Swiss mathematician, who was one of the 19 prominent mathematicians in the Bernoulli family.

**Binomial distribution.** Let $X$ be the number of successes among the $n$ independent Bernoulli trials. The distribution of $X$ is $Binomial(n, p)$, where $n$ is the pre-specified number of trials and $p$ is the probability of a success. We write $X \sim Bi(n, p)$. The pmf of $X$ is $p_X(x) = \mathbb{P}(X = x) = \binom{n}{x} p^x(1-p)^{n-x}$, $x = 0, 1, \ldots, n$. The mean of $X$ is $\mathbb{E}(X) = np$, and its variance is $\mathbb{V}ar(X) = np(1-p)$. The name of this distribution is derived from the binomial coefficient $\binom{n}{x}$,

which in turn, comes from the formula for the Newton's binomial $(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i}$. Here "binomial" means two terms, $a$ and $b$. This formula helps to show that the binomial probabilities sum up to one. Indeed, $\sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = (p+1-p)^n = 1$.

**Geometric distribution.** Let $X$ be the number of independent Bernoulli trials until the first success. The distribution of $X$ is $Geometric(p)$ where $p$ is the probability of a success. We write $X \sim Geom(p)$. The pmf of $X$ is $p_X(x) = \mathbb{P}(X = x) = p(1-p)^{x-1}, \quad x = 1, 2, 3, \ldots$. The mean of $X$ is $\mathbb{E}X = \dfrac{1}{p}$ and the variance is $\mathbb{V}ar(X) = \dfrac{1-p}{p^2}$.

**Note.** The name of the distribution comes from the geometric series. For example, we use the sum of infinite geometric series to show that the probabilities add up to one. We write $\sum_{x=1}^{\infty} p(1-p)^{x-1} = \dfrac{p}{1-p} \sum_{x=1}^{\infty} (1-p)^x = \dfrac{p}{1-p} \left( \dfrac{1}{1-(1-p)} - 1 \right) = \dfrac{p}{1-p} \cdot \dfrac{1-p}{p} = 1.$

**Note.** Another way to define a geometric distribution is to let $X$ be the number of failures until the first success. Then the pmf becomes $p_X(x) = \mathbb{P}(X = x) = p(1-p)^x, \quad x = 0, 1, 2, \ldots$. The mean is $\mathbb{E}X = \dfrac{1-p}{p}$ and the variance is $\mathbb{V}ar(X) = \dfrac{1-p}{p^2}$.

**Poisson distribution.** Let $X$ be the number of occurrences of a rare event. For example, the number of car accidents per month on a certain intersection, or the number of field mice per acre, or the number of typographical errors per page in a newspaper. The distribution of $X$ is $Poisson(\lambda)$ where the parameter $\lambda$ is the mean and variance of $X$ (that is, $\mathbb{E}X = \mathbb{V}ar(X) = \lambda$). We write $X \sim Poi(\lambda)$. The pmf of $X$ is $p_X(x) = \mathbb{P}(X = x) = \dfrac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, 2, \ldots$. In particular, $p(0) = \mathbb{P}(X = 0) = e^{-\lambda}, \quad p(1) = \mathbb{P}(X = 1) = \lambda e^{-\lambda}$, $p(2) = \mathbb{P}(X = 2) = \dfrac{\lambda^2}{2} e^{-\lambda}, \quad p(3) = \mathbb{P}(X = 3) = \dfrac{\lambda^3}{6} e^{-\lambda}$, etc. To show that these probabilities sum up to one, we can use the Taylor's expansion of the exponential function $e^y = \sum_{n=0}^{\infty} \dfrac{y^n}{n!}$. We write $\sum_{x=0}^{\infty} \dfrac{\lambda^x}{x!} e^{-\lambda} = e^{\lambda} e^{-\lambda} = 1$.

**Note.** The Poisson distribution is named after a French mathematician Simeon Poisson(1781 - 1840).

**Continuous Distributions**

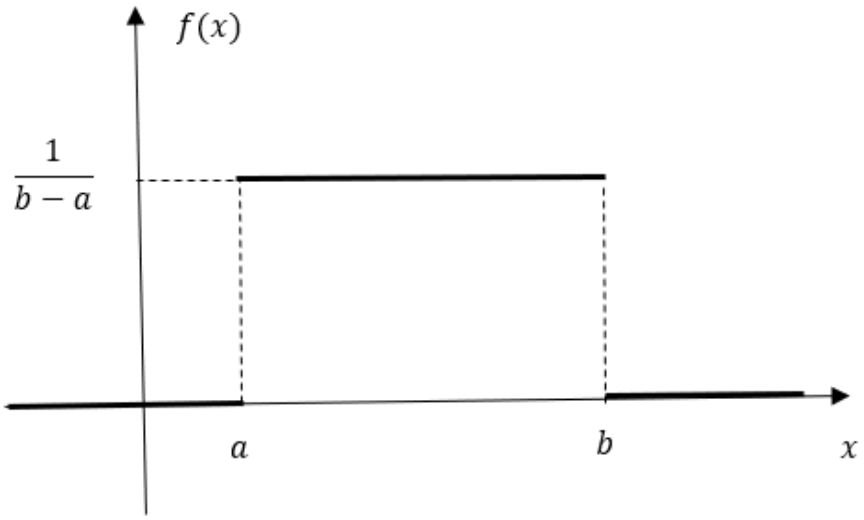
**Definition.** A continuous random variable $X$ is defined on an interval. The probability that $X$ is equal to any particular number is 0. To define the probability that $X$ falls between some values $a$ and $b$, the probability density function (pdf) is utilized. The pdf is defined as any function $f_X(x)$ with the properties: (i) $f(x) \geq 0$ for any $x, -\infty < x < \infty$, that is, the function is nowhere negative, and (ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$, that is, the total area under the curve is 1. Then we define the probability that $X$ is located between $a$ and $b$ with $b > a$ as $\mathbb{P}(a < X < b) = \int_a^b f(x)\,dx$. Note that technically speaking, it doesn't matter is we include or exclude the endpoints of the interval because $\mathbb{P}(X = a) = \mathbb{P}(X = b) = 0$. It means that $\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X < b) = \int_a^b f(x)\,dx$.

**Definition.** The cumulative distribution function (cdf) of $X$, $F_X(x), -\infty < x < \infty$, is defined as the area swept by the density function $f_X(x)$ up to $x$. That is, $F_X(x) = \int_{-\infty}^{x} f_X(u)\,du$.

When defining a continuous distribution, it is customary to specify both pdf and cdf, if the cdf has an explicit form. Specifying cdf eliminates the need to calculate integrals of the density when computing probabilities. We use the cdf as follows $\mathbb{P}(a < X < b) = F_X(b) - F_X(a)$. Indeed, $\mathbb{P}(a < X < b) = \int_a^b f_X(u)\,du = \int_{-\infty}^{b} f_X(u)\,du - \int_{-\infty}^{a} f_X(u)\,du = F_X(b) - F_X(a)$.

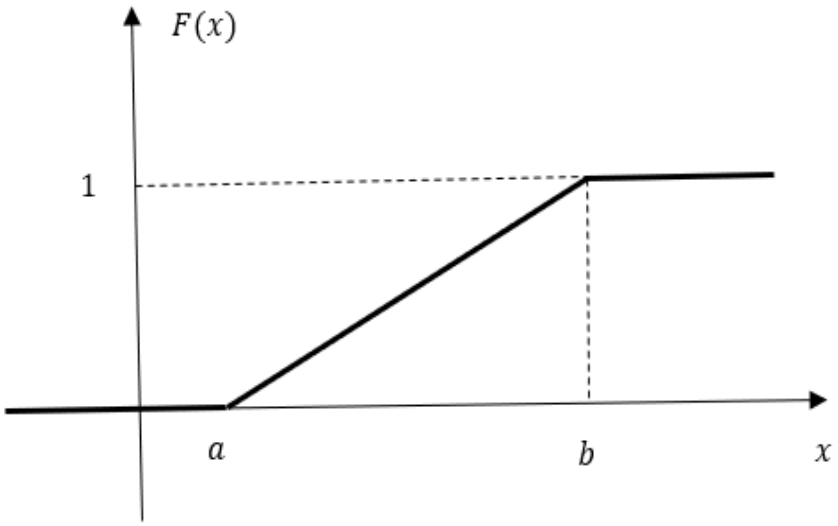
**Uniform distribution.** Suppose $X$ can assume any value in an interval $[a, b]$. The distribution of $X$ is $Uniform(a, b)$. We write $X \sim Unif(a, b)$. The pdf of $X$ is

$$f_X(x) = \begin{cases} 0, & x < a, \\ \frac{1}{b-a}, & a \leq x \leq b, \\ 0, & x > b. \end{cases}$$




The cdf of $X$ is

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du = \begin{cases} 0, & x < a, \\ \int_a^x \frac{1}{b-a}\,du = \frac{x-a}{b-a}, & a \le x \le b, \\ 1, & x > b. \end{cases}$$




The mean and variance of $X$ are $\mathbb{E}X = \frac{a+b}{2}$ and $\mathbb{V}ar(X) = \frac{(b-a)^2}{12}$. Note that the mean is the middle of the interval $[a, b]$, and the variance is the length of the interval $b - a$ squared divided by 12.

The **standard uniform distribution** is uniform distribution on the interval $[0,1]$. We write $U \sim Unif(0,1)$. The pdf of $U$ is $f_U(u) = 1$ if $0 \le u \le 1$ and 0 otherwise. The cdf of $U$ is

$$F_U(u) = \begin{cases} 0, & u < 0, \\ u, & 0 \le u \le 1, \\ 1, & u > 1. \end{cases}$$

The mean of $U$ is $\mathbb{E}U = 1/2$ and variance is $\mathbb{V}ar(U) = \mathbb{E}U^2 - (\mathbb{E}U)^2 = \int_0^1 u^2\,du - (1/2)^2 = 1/3 - 1/4 = 1/12$. This explains the 12 in the denominator (the second moment minus the square of the first moment is $1/3 - 1/4 = 1/12$).

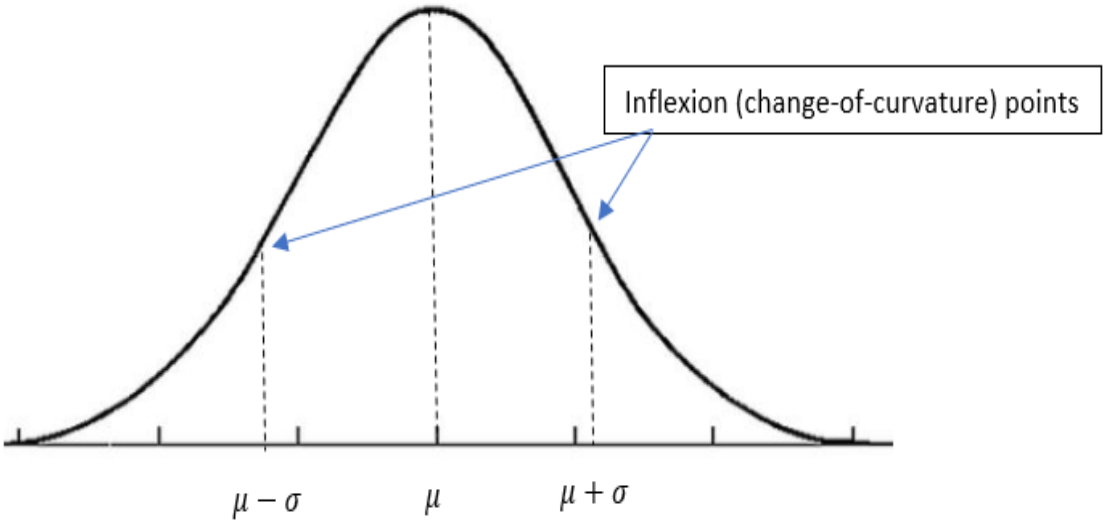
**Exponential distribution.** A continuous random variable $X$ has an exponential distribution with mean $\beta$ (written $X \sim Exp(mean = \beta)$) if the pdf of $X$ is

$$f_X(x) = \frac{1}{\beta}\, e^{-x/\beta}, \quad x > 0,\ \beta > 0.$$

The cdf of $X$ is $F_X(x) = 1 - e^{-x/\beta},\ x > 0$. The mean and variance of $X$ are $\mathbb{E}X = \beta$, and $\mathbb{V}ar(X) = \beta^2$. Note that if the parameter $\beta$ is the mean, then in the pdf and cdf we divide by $\beta$. Sometimes the pdf of an exponential distribution is defined as $f_X(x) = \beta\, e^{-x\beta},\ x > 0$. Then $\mathbb{E}X = 1/\beta$ and $\mathbb{V}ar(X) = 1/\beta^2$.

**Normal distribution.** A continuous random variable $X$ has a normal distribution with mean $\mu$ and variance $\sigma^2$ (written $X \sim N(\mu, \sigma^2)$) if the pdf of $X$ is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, \exp\Big\{ -\frac{(x-\mu)^2}{2\sigma^2} \Big\}, \quad -\infty < x < \infty.$$




The cdf of $X$ doesn't have a closed form.

The **standard normal distribution** is a normal distribution with mean $\mu = 0$, and variance $\sigma^2 = 1$. As a rule, a standard normal random variable is denoted by $Z$. We write $Z \sim N(0,1)$. The pdf of $Z$ is $f_Z(z) = \dfrac{1}{\sqrt{2\pi}} \exp\Big\{-\dfrac{z^2}{2}\Big\}$, $-\infty < z < \infty$.

**Note.** To show that the density integrates to 1, we need to show that $I = \displaystyle\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = 1$. To this end, we write

$$I^2 = \Big(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx\Big)\Big(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy\Big) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\, dxdy$$

$$= \{\text{polar coordinates} : r^2 = x^2+y^2, dxdy = rdrd\theta\} = \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\infty} e^{-r^2/2}\, rdrd\theta$$

$$= \frac{1}{2\pi}\int_0^{2\pi} d\theta \int_0^{\infty} e^{-r^2/2}\, d(r^2/2) = 1.$$

Further, the cdf of a standard normal distribution is traditionally denoted by

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\Big\{-\frac{u^2}{2}\Big\}\, du,$$

and it is tabulated for various values of $z$.

Even though the values are tabulated only for the standard normal random variable $Z$, we can use the table to compute the cdf of any normally distributed random variable $X \sim N(\mu, \sigma^2)$ if we use the relation $X = \mu + Z\sigma$, or, equivalently, $Z = (X - \mu)/\sigma$. Putting it in words, $Z$ represents how many standard deviations ($\sigma$s) the random variable $X$ is above or below its mean $\mu$. For example, suppose $X \sim N(1, 4)$. Then $\mathbb{P}(X < 2) = \mathbb{P}\Big(\dfrac{X - \mu}{\sigma} < \dfrac{2 - 1}{2}\Big) = \mathbb{P}(Z < 0.5) = 0.6915$.

**Note.** Normal distribution is sometimes called **Gaussian distribution**. It was introduced by Johann Carl Friedrich Gauss (1777 – 1855) who was a German mathematician.

### The Central Limit Theorem

Suppose we have a sequence of independent random variables $X_1, X_2, \ldots$ that have the same distribution with mean $\mathbb{E}X_1 = \mu$ and variance $\mathbb{V}ar(X_1) = \sigma^2$. Define $\bar{X}_n = (X_1 + \cdots + X_n)/n$. Note that $\mathbb{E}\bar{X}_n = \mu$ and $\mathbb{V}ar(\bar{X}_n) = \sigma^2/n$. The **Central Limit Theorem** (CLT) states that for large $n$ (in practice, $n \geq 30$), $\dfrac{\bar{X}_n - \mathbb{E}\bar{X}_n}{\sqrt{\mathbb{V}ar(\bar{X}_n)}} = \dfrac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ has approximately $N(0, 1)$ distribution.

**Example.** Twenty-ounce Coke bottles contain on average 20 oz of liquid with the standard deviation of 0.3 oz. Suppose we want to compute the probability that in a random sample of 81 Coke bottles the sample mean is above 20.05 oz. By the CLT, $\bar{X}_{81} \overset{approx.}{\sim} N\big(20, (0.3)^2\big)$. Hence, $\mathbb{P}(\bar{X}_{81} > 20.05) = \mathbb{P}\big(Z > (20.05 - 20)/(0.3/\sqrt{81})\big) = \mathbb{P}(Z > 1.5) = 0.0668$. □

### Order Statistics

**Definition.** Suppose we have $n$ observations $X_1, \ldots, X_n$. Denote by $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ the ordered set. For any $i$, $i = 1, \ldots, n$, $X_{(i)}$ is called the **$i$-th order statistic**. Note that $X_{(1)}$ is the minimum, whereas $X_{(n)}$ denotes the maximum.

**Proposition.** Suppose $X_1, \ldots, X_n$ are independent and identically distributed (iid) random variables with common pdf $f(x)$ and cdf $F(x)$. The pdf of the $i$-th order statistic has the form

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!}\big[F(x)\big]^{i-1} f(x)\big[1 - F(x)\big]^{n-i}.$$

PROOF: If the $i$-th order statistic is "equal" to $x$ (contributing $f(x)$), then $i - 1$ observations necessarily lie below $x$ (contributing $\big[F(x)\big]^{i-1}$) , and the other $n - i$ lie above $x$ (contributing $\big[1 - F(x)\big]^{n-i}$). Finally, the multiplicative factor is the number of ways to choose $i - 1$ observations to lie below $x$, and $n - i$ to exceed $x$. □

**Distribution of Maximum.** If we let $i = n$ in the above proposition, we obtain the pdf of the maximum of $n$ iid observations,

$$f_{X_{(n)}}(x) = \frac{n!}{(n-1)!(n-n)!}\big[F(x)\big]^{n-1} f(x)\,\big[1-F(x)\big]^{n-n} = n\,f(x)\,\big[F(x)\big]^{n-1}.$$

This is intuitive, since the pdf of $X_{(n)}$ can also be obtained by the following reasoning:

$$F_{X_{(n)}}(x) = \mathbb{P}(X_{(n)} \le x) = \mathbb{P}(X_1 \le x,\, X_2 \le x, \ldots, X_n \le x) = \big[F(x)\big]^n,$$

and, thus, the pdf is $f_{X_{(n)}}(x) = F'_{X_{(n)}}(x) = n\,f(x)\,\big[F(x)\big]^{n-1}$.

**Distribution of Minimum.** In the formula for the pdf of the $i$-th order statistic we let $i = 1$ to obtain that $f_{X_{(1)}}(x) = \dfrac{n!}{(1-1)!(n-1)!}\big[F(x)\big]^{1-1} f(x)\,\big[1-F(x)\big]^{n-1} = n\,f(x)\,\big[1-F(x)\big]^{n-1}$. We can also find the pdf of the minimum as follows:

$$1 - F_{X_{(1)}}(x) = \mathbb{P}(X_{(1)} \ge x) = \mathbb{P}(X_1 \ge x,\, X_2 \ge x, \ldots, X_n \ge x) = \big[1-F(x)\big]^n,$$

therefore, $F_{X_{(1)}}(x) = 1 - \big[1-F(x)\big]^n$, and $f_{X_{(n)}}(x) = F'_{X_{(n)}}(x) = n\,f(x)\,\big[1-F(x)\big]^{n-1}$.

**Example.** Let $X_1, \ldots, X_n$ be independent exponential random variables with mean $\beta$. The pdf is $f(x) = \dfrac{1}{\beta}\,\exp\{-\dfrac{x}{\beta}\}$, and the cdf is $F(x) = 1 - \exp\{-\dfrac{x}{\beta}\},\ x > 0,\ \beta > 0$. Therefore,

(a) the $i$-th order statistic has the pdf

$$\begin{aligned} f_{X_{(i)}}(x) &= \frac{n!}{(i-1)!(n-i)!}\,\big[1-\exp\{-\frac{x}{\beta}\}\big]^{i-1}\,\frac{1}{\beta}\,\exp\{-\frac{x}{\beta}\}\,\big[\exp\{-\frac{x}{\beta}\}\big]^{n-i} \\ &= \frac{n!}{(i-1)!(n-i)!}\,\frac{1}{\beta}\,\exp\{-\frac{x\,(n-i+1)}{\beta}\}\,\big[1-\exp\{-\frac{x}{\beta}\}\big]^{i-1}. \end{aligned}$$

(b) The pdf of the maximum is derived by letting $i = n$. We have $f_{X_{(n)}}(x) = \dfrac{n}{\beta}\,\exp\{-\dfrac{x}{\beta}\}\,\big[1-\exp\{-\dfrac{x}{\beta}\}\big]^{n-1}$. We can also notice that the cdf of the maximum is $F(x) = \big(1-\exp\{-\dfrac{x}{\beta}\}\big)^n$, which can be obtained by either integrating the density or arguing that all $n$ observations must not exceed $x$, if the maximum doesn't exceed $x$.

(c) In particular, for $i = 1$, the pdf of the minimum is $f_{X_{(1)}}(x) = \dfrac{n}{\beta}\,\exp\{-\dfrac{n}{\beta}\,x\}$, that is, $X_{(1)}$ has an exponential distribution with mean $\dfrac{\beta}{n}$. $\square$

## NEW MATERIAL: MAXIMUM LIKELIHOOD ESTIMATION METHOD

**Definition.** Suppose $X_1, \ldots, X_n$ are iid random variables with a common pmf (discrete case) or pdf (continuous case) $f(x;\theta)$. The *likelihood function* is a function of the unknown parameter $\theta$ that is given by

$$L(\theta) = L(\theta \,|\, X_1, \ldots, X_n) = \prod_{i=1}^{n} f(X_i;\theta).$$

**Note.** The likelihood function represents the probability to observe the data points that have been observed, namely, $X_1, \ldots, X_n$.

**Definition.** An estimator $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ is called the *maximum likelihood estimator* (*MLE*) of $\theta$ if it maximizes the likelihood function $L(\theta)$, that is, it maximizes the probability to observe the data points that have been observed. To find an MLE of $\theta$, one needs to differentiate $L(\theta)$ with respect to $\theta$, set the derivative equal to zero, and solve for $\theta$. Technically speaking, one has to show also that the second derivative of $L(\theta)$ with respect to $\theta$ is negative at the point where the first derivative is zero, so that the attained extremum is a maximum not minimum, but for all basic distributions the extremum is in fact a maximum, so there is no need to verify the condition for the second derivative.

**Example (Bernoulli distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} Ber(p)$. The likelihood function is

$$L(p \,|\, X_1, \ldots, X_n) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}.$$

It is easier to work with the **log-likelihood function**, the natural logarithm of the likelihood function,

$$\ln L(p \,|\, X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i \, \ln p + (n - \sum_{i=1}^{n} X_i) \, \ln(1-p).$$

**Note.** Since the logarithm is a strictly increasing function, the maximum of the log-likelihood function is attained at the same point where the maximum of the likelihood function itself is attained. Put mathematically, $\dfrac{d \ln L(\theta)}{d\theta} = \dfrac{L'(\theta)}{L(\theta)} = 0$ if and only if $L'(\theta) = 0$.

Further, to maximize the log-likelihood function, we equate to zero the first partial derivative of $\ln L(p \,|\, X_1, \ldots, X_n)$ with respect to $p$, and solve for $p$. We obtain

$$0 = \frac{d \ln L(p \,|\, X_1, \ldots, X_n)}{dp} = \frac{\sum_{i=1}^n X_i}{p} - \frac{n - \sum_{i=1}^n X_i}{1-p}.$$

Thus, $\widehat{p}$, the maximum likelihood estimator of $p$, satisfies the equation

$$\frac{\sum_{i=1}^n X_i}{\widehat{p}} = \frac{n - \sum_{i=1}^n X_i}{1-\widehat{p}},$$

from where $\sum_{i=1}^{n} X_i - \widehat{p}\sum_{i=1}^{n} X_i = n\widehat{p} - \widehat{p}\sum_{i=1}^{n} X_i$, or $\widehat{p} = \sum_{i=1}^{n} X_i/n = \bar{X}$. The MLE $\widehat{p} = \bar{X}$ represents the proportion of successes among $n$ observations, and is an intuitive estimator of $p$, the probability of a success. For instance, if we observe a sequence 0, 1, 1, 1, 0, 0, 1, 0, 1, the MLE of $p$ is $\widehat{p}$ =proportion of ones= 5/9. $\square$

**Example (geometric distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} Geom(p)$ with pmf $p(x) = p(1-p)^{x-1}$, $x = 1, 2, \ldots$. The likelihood function has the form

$$L(p \,|\, X_1, \ldots, X_n) = \prod_{i=1}^{n} p\,(1-p)^{X_i - 1} = p^n\,(1-p)^{\sum_{i=1}^{n} X_i - n}.$$

The log-likelihood function is

$$\ln L(p \,|\, X_1, \ldots, X_n) = n\,\ln p + \big(\sum_{i=1}^{n} X_i - n\big)\,\ln(1-p).$$

The MLE $\widehat{p}$ solves the equation

$$0 = \frac{d \ln L(p \,|\, X_1, \ldots, X_n)}{dp}\Big|_{p=\widehat{p}} = \frac{n}{\widehat{p}} - \frac{\sum_{i=1}^{n} X_i - n}{1-\widehat{p}},$$

and so,

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

Since the mean of $X_i$'s is equal to $1/p$, the MLE is an estimator of $p$ derived from estimating the mean by the sample mean $\bar{X}$. $\square$

**Example (Poisson distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} Poi(\lambda)$. The likelihood function is

$$L(\lambda \,|\, X_1, \ldots, X_n) = \prod_{i=1}^{n} \frac{\lambda^{X_i}\,\exp\{-\lambda\}}{X_i!} = \Big[\prod_{i=1}^{n} \frac{1}{X_i!}\Big]\,\lambda^{\sum_{i=1}^{n} X_i}\,\exp\{-n\lambda\},$$

and the log-likelihood function takes the form

$$\ln L(\lambda \,|\, X_1, \ldots, X_n) = \ln\Big[\prod_{i=1}^{n} \frac{1}{X_i!}\Big] + \sum_{i=1}^{n} X_i\,\ln\lambda - n\lambda.$$

The MLE $\widehat{\lambda}$ is the solution of the equation

$$0 = \frac{d \ln L(\lambda \,|\, X_1, \ldots, X_n)}{d\lambda}\Big|_{\lambda=\widehat{\lambda}} = \frac{\sum_{i=1}^{n} X_i}{\widehat{\lambda}} - n.$$

Hence,

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

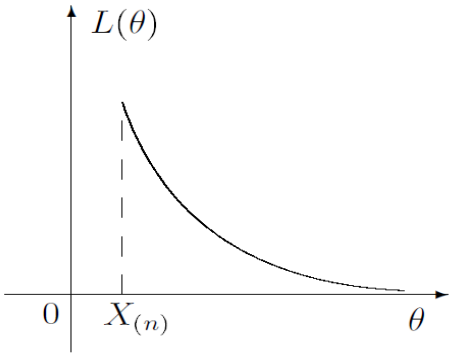
Indeed, it is intuitive to estimate the mean $\lambda$ by the sample mean$\bar{X}$. For example, let 4, 4, 2, 0, 3, 1, 1, 5, 3, 1 come from a Poisson distribution with parameter $\lambda$. The MLE of $\lambda$ is $\widehat{\lambda} = \bar{X} = 24/10 = 2.4$. □

**Example (uniform distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} Unif(0, \theta)$. The likelihood function is derived as

$$L(\theta \,|\, X_1, \ldots, X_n) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{I}\{0 \le X_i \le \theta\} = \frac{1}{\theta^n} \mathbb{I}\{0 \le X_{(n)} \le \theta\}.$$

Here $\mathbb{I}\{A\}$ denotes the indicator function of an event $A$, that is, it is equal to 1 if $A$ occurs, and 0, otherwise. The last equality is justified by noticing that the events $\{0 \le X_i \le \theta\}$ occur simultaneously for all $i = 1, \ldots, n$, if and only if the event $\{0 \le X_{(n)} \le \theta\}$ occurs.

Next, we plot the likelihood function $L(\theta) = L(\theta \,|\, X_1, \ldots, X_n) = 1/\theta^n, \;\; \theta \ge X_{(n)}$, against $\theta$ to see where it attains the maximum value.




As seen on the graph, the maximum is attained at $X_{(n)}$, thus $\widehat{\theta} = X_{(n)}$ is the MLE of $\theta$. On intuitive level, if $X_1, \ldots, X_n$ are observed, and we know that each of them doesn't exceed $\theta$, then our best guess about the value of $\theta$ is the maximum of all the observations.

For example, suppose the observations are 0.156, 0.324, 0.011, 0.896, 0.376, 0.423, 0.799, and 0.206, and we know that they come from a uniform distribution of the interval $[0, \theta]$. The MLE of $\theta$ is the maximum of these observations which is $\widehat{\theta} = 0.896$. □

**Example (exponential distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} Exp$ with mean $\beta$. The likelihood function is written as

$$L(\beta \,|\, X_1, \ldots, X_n) = \prod_{i=1}^{n} \frac{1}{\beta} \exp\{-X_i/\beta\} = \frac{1}{\beta^n} \exp\{-\sum_{i=1}^{n} X_i/\beta\},$$

and the log-likelihood function takes the form

$$\ln L(\beta \,|\, X_1, \ldots, X_n) = -n \ln \beta - \frac{\sum_{i=1}^{n} X_i}{\beta}.$$

The maximum likelihood estimator of $\beta$ satisfies the equation

$$0 = \frac{d \ln L(\beta \,|\, X_1, \ldots, X_n)}{d\beta}\Big|_{\beta=\widehat{\beta}} = -\frac{n}{\widehat{\beta}} + \frac{\sum_{i=1}^{n} X_i}{\widehat{\beta}^2}.$$

From here,

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

We see that it is only reasonable to estimate the mean $\beta$ by the sample mean $\bar{X}$. $\square$

**Example (shifted exponential distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta) = \exp\{-(x-\theta)\}, \;\; x > \theta$. This distribution is called **shifted exponential distribution**. It is an exponential distribution with mean 1 shifted by $\theta$. Its mean is equal $1+\theta$ and variance is 1. We need to find the MLE of $\theta$. Note that since the range of $x$ depends on $\theta$, the MLE is an order statistics. The log-likelihood function has the form

$$L(\theta) = \prod_{i=1}^{n} \exp\{-(X_i-\theta)\}\, \mathbb{I}\{X_i \geq \theta\} \;=\; \exp\{-\sum_{i=1}^{n}(X_i-\theta)\}\, \mathbb{I}\{X_{(1)} \geq \theta\} \;=\; \exp\{-n(\bar{X}-\theta)\}\, \mathbb{I}\{X_{(1)} \geq \theta\}.$$

This is an exponentially increasing function of $\theta$ which reaches its maximum in the rightmost point $X_{(1)}$. Thus, the MLE of $\theta$ is the minimum of the observations. $\square$

**Note.** Like in the case of uniform and shifted exponential distributions, if the range of $x$ depends on $\theta$, the MLE of $\theta$ is always an extreme order statistic (minimum or maximum).

**Example (normal distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ where both, $\mu$ and $\sigma$ are unknown. First, we obtain the likelihood function. We write

$$L(\mu, \sigma^2 \,|\, X_1, \ldots, X_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{ -\frac{(X_i-\mu)^2}{2\sigma^2} \Big\}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big\{ - \frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^2} \Big\}.$$

Next, we find the log-likelihood function as

$$\ln L(\mu, \sigma^2 \,|\, X_1, \ldots, X_n) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^2}.$$

The maximum likelihood estimators $\widehat{\mu}$ and $\widehat{\sigma}^2$ are solutions of the system of two equations

$$\begin{cases} 0 = \dfrac{\partial \ln L(\mu, \sigma^2 \,|\, X_1, \ldots, X_n)}{\partial \mu}\Big|_{\substack{\mu=\widehat{\mu},\\ \sigma^2=\widehat{\sigma}^2}} = \dfrac{\sum_{i=1}^n (X_i - \widehat{\mu})}{\widehat{\sigma}^2}, \\ 0 = \dfrac{\partial \ln L(\mu, \sigma^2 \,|\, X_1, \ldots, X_n)}{\partial \sigma^2}\Big|_{\substack{\mu=\widehat{\mu},\\ \sigma^2=\widehat{\sigma}^2}} = -\dfrac{n}{2\widehat{\sigma}^2} + \dfrac{\sum_{i=1}^n (X_i - \widehat{\mu})^2}{2\widehat{\sigma}^4}, \end{cases}$$

so

$$\widehat{\mu} = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}, \quad \text{and} \quad \widehat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}.$$

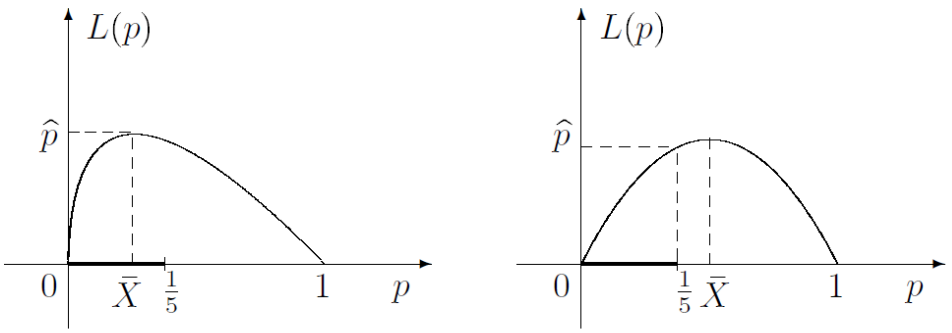
Since $\mu$ is the mean of the normal distribution, the estimator is indeed intuitive. The variance is estimated by the average squared distance between each observation and the sample mean, which is a natural measure of spread. $\square$

**Example (Bernoulli distribution with a constraint).** Let $X_1, \ldots, X_n \overset{iid}{\sim} Ber(p)$, where $p$ is constrained by the condition $p \le 1/5$. In a previous example we have shown that the maximum of the likelihood function

$$L(p) = L(p \,|\, X_1, \ldots, X_n) = p^{\sum_{i=1}^n X_i} (1-p)^{n - \sum_{i=1}^n X_i}$$

is attained when $p = \bar{X}$.

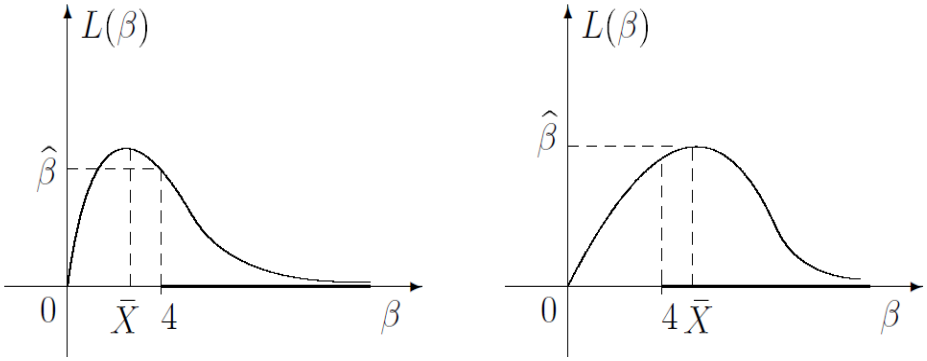
We will plot this likelihood function against values of $p$ when $\bar{X}$ is on either side of $1/5$ to see where the maximum of this function is attained on $[0, 1/5]$.



From the graphs, if $0 \le \bar{X} \le 1/5$, then the maximum of $L(p)$ on the interval $0 \le p \le 1/5$ is attained at $\bar{X}$, whereas when $\bar{X} > 1/5$, then the maximum of the likelihood function on this interval is attained at $1/5$. Thus, the MLE of $p$ is

$$\widehat{p} = \begin{cases} \bar{X}, & \text{if } 0 \le \bar{X} \le 1/5, \\ 1/5, & \text{if } \bar{X} > 1/5. \end{cases} \qquad \square$$

**Example(exponential distribution with a constraint).** Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta) = \frac{1}{\beta} e^{-x/\beta}$, $x > 0$, with an additional constraint that $\beta > 4$. We know from a previous example that in the general case of $\beta > 0$, the likelihood function $L(\beta) = \dfrac{1}{\beta^n} \exp\{-\sum_{i=1}^{n} X_i/\beta\}$ attains its maximum at $\widehat{\beta} = \bar{X}$. In this case, the values of $\beta$ are bounded from below by 4. The two graphs present two possible scenarios: when $0 \le \bar{X} < 4$ and when $\bar{X} \ge 4$.




As seen on the graphs, the maximum of the likelihood function is attained on $[4, \infty)$ at $\widehat{\beta} = 4$ if $0 \le \bar{X} < 4$, and at $\widehat{\beta} = \bar{X}$, if $\bar{X} \ge 4$. $\square$

**Example (discrete distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$ where the pmf $f(x;\theta)$ is given by the table:

| | $x$ | | |
|---|---|---|---|
| $\theta$ | 1 | 2 | 4 |
| 0 | 1/4 | 1/2 | 1/4 |
| 1/3 | 1/2 | 0 | 1/2 |
| 1/4 | 3/5 | 1/5 | 1/5 |

Suppose the observations are $X_1 = 1$, $X_2 = 4$, and $X_3 = 2$. We need to find the MLE of $\theta$. The likelihood function is calculated as

$$L(\theta; X_1, X_2, X_3) = f(1;\theta)f(4;\theta)f(2;\theta) = \begin{cases} (1/4)(1/4)(1/2) = 0.03125, & \text{if } \theta = 0, \\ (1/2)(1/2)(0) = 0, & \text{if } \theta = 1/3, \\ (3/5)(1/5)(1/5) = 0.024, & \text{if } \theta = 1/4. \end{cases}$$

The largest value of the likelihood function is 0.03125 and corresponds to the MLE $\widehat{\theta} = 0$. $\square$.

**Example (discrete distribution).** Suppose $X$ has pmf given in the table below.

| $x$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $p(x)$ | $3\theta/5$ | $2\theta/5$ | $2(1-\theta)/5$ | $3(1-\theta)/5$ |

Suppose we observe 0, 2, 1, 3, 0, 0, 3, 1, 2, 2, 0, 1, and 1. The likelihood function can be written as

$$L(\theta) = \Big(\frac{3\theta}{5}\Big)^4 \Big(\frac{2\theta}{5}\Big)^4 \Big(\frac{2(1-\theta)}{5}\Big)^3 \Big(\frac{3(1-\theta)}{5}\Big)^2.$$

The log-likelihood function has the form $\ln L(\theta) = 4\ln\theta + 4\ln\theta + 3\ln(1-\theta) + 2\ln(1-\theta) + constant$. We take the first derivative and set it equal to zero. We obtain $L'(\theta) = \dfrac{4}{\theta} + \dfrac{4}{\theta} - \dfrac{3}{1-\theta} - \dfrac{2}{1-\theta} = 0$, or, equivalently, $\dfrac{8}{\theta} = \dfrac{5}{1-\theta}$. Solving for $\theta$, we get $\widehat{\theta} = 8/13$. $\square$

**Theorem (Functional Invariance of MLE).** Suppose $X_1, \ldots, X_n \overset{iid}{\sim}$ pmf or pdf $f(x;\theta)$. Let $g$ be some continuous function and let $\widehat{\theta}$ denote the MLE of $\theta$. Then the MLE of $g(\theta)$ can be computed as $\widehat{g}(\theta) = g(\widehat{\theta})$.

**Example (Bernoulli distribution - function of parameter).** Let $X_1, \ldots, X_n \overset{iid}{\sim} Ber(p)$. Suppose we need to find the MLE of the variance $\mathbb{V}ar(X_1) = p(1-p)$. We know that the MLE of $p$ is $\widehat{p} = \bar{X}$, and by the functional invariance of MLE theorem, the MLE of the variance of $X_1$ is $\widehat{\mathbb{V}ar}(X_1) = \widehat{p}(1-\widehat{p}) = \bar{X}(1-\bar{X})$. $\square$

**Example (Poisson distribution - function of parameter).** Let $X_1, \ldots, X_n \overset{iid}{\sim} X \sim Poi(\lambda)$. Suppose we need to find the MLE of $\mathbb{P}(X_1 = 0) = \exp\{-\lambda\}$. We know that the MLE of $\lambda$ is $\widehat{\lambda} = \bar{X}$. So, by the invariance of MLE theorem, the MLE of $\mathbb{P}(X_1 = 0)$ is $\widehat{\mathbb{P}}(X_1 = 0) = \exp\{-\bar{X}\}$. Similarly, the MLE of, say, $\mathbb{P}(X_1 = 2) = \dfrac{\lambda^2}{2}\exp\{-\lambda\}$ is $\widehat{\mathbb{P}}(X_1 = 2) = \dfrac{\bar{X}^2}{2}\exp\{-\bar{X}\}$. $\square$

**Example (uniform distribution - function of parameter).** Let $X_1, \ldots, X_n \overset{iid}{\sim} Unif(0,\theta)$. Suppose we need to find the MLE of $\mathbb{V}ar(X_1) = \theta^2/12$. As we already know that the MLE of $\theta$

is $\widehat{\theta} = X_{(n)}$. Applying the theorem on functional invariance of MLE, we get that the MLE of the variance is $\widehat{\mathbb{V}ar}(X_1) = X_{(n)}^2/12$. $\square$

**Example (variation of Bernoulli distribution - function of parameter).** Suppose $X_1, \ldots, X_n$ are iid with the pmf $p(0) = e^{-\theta}$ and $p(1) = 1 - e^{-\theta}$. Assume that we need to find the MLE of $\theta$. Denote by $p = 1 - e^{-\theta}$. Here $p$ is the probability of a success for a Bernoulli distribution and we already know that the MLE of $p$ is $\widehat{p} = \bar{X}$. Now we solve for $\theta$ the equation $p = 1 - e^{-\theta}$. We get $\theta = -\ln(1-p)$, and therefore, by the invariance principle, the MLE $\widehat{\theta} = -\ln(1-\widehat{p}) = -\ln(1-\bar{X})$. $\square$

## METHOD OF MOMENTS ESTIMATION

**Definition.** Suppose $X_1, \ldots, X_n$ are iid random variables with a common distribution that depends on $k$ parameters $\theta_1, \ldots, \theta_k$. The **method of moments** (MM) estimators of the parameters solve the system of $k$ equations:

$$\begin{cases} \mathbb{E}(X_1) = \dfrac{\sum_{i=1}^n X_i}{n} = \bar{X}, \\ \mathbb{E}(X_1^2) = \dfrac{\sum_{i=1}^n X_i^2}{n}, \\ \mathbb{E}(X_1^3) = \dfrac{\sum_{i=1}^n X_i^3}{n}, \\ \cdots \\ \mathbb{E}(X_1^k) = \dfrac{\sum_{i=1}^n X_i^k}{n}. \end{cases}$$

That is, in each equation the theoretical moment is equated to the corresponding empirical moment.

**Example (Bernoulli distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} Ber(p)$. To find the MM estimator of $p$, we equate the theoretical and empirical first moments. We have

$$\mathbb{E}(X_1) = p = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}.$$

The solution is $\widehat{p} = \bar{X}$, and, thus, the MM estimator coincides with the MLE for $p$. $\square$

**Example (geometric distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} Geom(p)$. The MM estimator for $p$ satisfies

$$\mathbb{E}(X_1) = \frac{1}{p} = \bar{X}.$$

Hence, $\widehat{p} = 1/\bar{X}$, which is the same as the MLE for $p$. $\square$

**Example (Poisson distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} Poi(\lambda)$. The MM estimator for $\lambda$ is the solution of the equation

$$\mathbb{E}(X_1) = \lambda = \bar{X},$$

and so, $\widehat{\lambda} = \bar{X}$. It is the same as the MLE. $\square$

**Example (uniform distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} Unif(0, \theta)$. To find the MM estimator for $\theta$ we write

$$\mathbb{E}(X_1) = \frac{\theta}{2} = \bar{X},$$

thus, $\widehat{\theta} = 2\bar{X}$. This estimator is not the same as $X_{(n)}$, the MLE of $\theta$. Moreover, for some observations, $2\bar{X}$ is smaller than $X_{(n)}$, and hence, the MM estimator doesn't always make sense. For example, if $X_1 = 1, X_2 = 1, X_3 = 2$, and $X_4 = 8$. Then $2\bar{X} = 6$, whereas $X_{(4)} = 8$, so we have an observation that exceeds our MM estimate of $\theta$. $\square$

**Example (exponential distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} Exp$ with mean $\beta$. The MM estimator for $\beta$ is the solution of the equation $\mathbb{E}(X_1) = \beta = \bar{X}$, thus, $\widehat{\beta} = \bar{X}$, and is equal to the MLE. $\square$

**Example (normal distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. To find the MM estimators of $\mu$ and $\sigma^2$, we equate the first and second theoretical and empirical moments, respectively:

$$\begin{cases} \mathbb{E}(X_1) = \mu = \dfrac{\sum_{i=1}^n X_i}{n} = \bar{X}, \\ \mathbb{E}(X_1^2) = \sigma^2 + \mu^2 = \dfrac{\sum_{i=1}^n X_i^2}{n}. \end{cases}$$

The solution of this system is $\widehat{\mu} = \bar{X}$, and $\widehat{\sigma}^2 = \dfrac{\sum_{i=1}^n X_i^2}{n} - \bar{X}^2 = \dfrac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$. Note that the MM estimators of $\mu$ and $\sigma^2$ coincide with the corresponding MLEs. $\square$

## PROPERTIES OF ESTIMATORS: UNBIASEDNESS AND CONSISTENCY

Let $X_1, \ldots, X_n \overset{iid}{\sim}$ pmf or pdf $f(x; \theta)$. We can look at different estimators of $\theta$, for example, $\widehat{\theta}_{MLE}, \widehat{\theta}_{MM}, \widehat{\theta}_1 = X_1, \widehat{\theta}_2 = (X_1 + X_2)/2$, and $\widehat{\theta}_3 = 1$. Which one is better? Good estimators should have two properties: (i) unbiasedness, and (ii) consistency.

**Definition.** An estimator $\widehat{\theta}$ is called **unbiased** if $\mathbb{E}(\widehat{\theta}) = \theta$. An estimator that is not unbiased is called **biased**. Simply put, unbiasedness means that $\widehat{\theta}$, on average, estimates $\theta$. It doesn't

systematically underestimates or overestimates it.

**Definition.** An estimator $\widehat{\theta}$ is a function of $X_1, \ldots, X_n$. We will index this estimator by $n$ and write $\widehat{\theta}_n$. This estimator is a **consistent** estimator of $\theta$ if it is unbiased and $\mathbb{V}ar(\widehat{\theta}_n) \to 0$ as $n \to \infty$.

Note that a consistent estimator is necessarily unbiased.

**Example (Bernoulli distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} Ber(p)$, and consider $\widehat{p} = \bar{X}$, the MLE and MM estimator of $p$. This estimator is unbiased because $\mathbb{E}(\widehat{p}) = \mathbb{E}(\bar{X}) = \mathbb{E}(X_1) = p$. The estimator is consistent since $\mathbb{V}ar(\bar{X}) = \mathbb{V}ar(X_1)/n = p(1-p)/n \to 0$ as $n \to \infty$. $\square$.

**Note:** For any distribution, an estimator $\bar{X}$ is an unbiased estimator of the mean since $\mathbb{E}(\bar{X}) = \mathbb{E}(X_1)$. And also, this estimator is consistent because $\mathbb{V}ar(\bar{X}) = \mathbb{V}ar(X_1)/n \to 0$ as $n \to \infty$.

**Example (geometric distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} Geom(p)$. We will show that the MLE and MM estimator $\widehat{p} = 1/\bar{X}$ is a biased estimator of $p$. The sum $\sum_{i=1}^n X_i$ of $n$ independent $Geom(p)$ random variables has a negative binomial distribution with parameters $n$ and $p$. Its pmf can be written as

$$P\big(\sum_{i=1}^n X_i = x\big) = \binom{x-1}{n-1} p^n (1-p)^{x-n}, \quad x = n, n+1, \ldots.$$

So, we write

$$\mathbb{E}(\widehat{p}) = \mathbb{E}\big(\frac{1}{\bar{X}}\big) = \mathbb{E}\Big(\frac{n}{\sum_{i=1}^n X_i}\Big) = \sum_{x=n}^{\infty} \frac{n}{x} \binom{x-1}{n-1} p^n (1-p)^{x-n} \neq p.$$

Thus, the estimator is biased, and so it is not consistent. $\square$

**Example (Poisson distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} Poi(\lambda)$. The MLE and MM estimator $\widehat{\lambda} = \bar{X}$ is an unbiased and consistent estimator of $\lambda$ since, as pointed out above $\mathbb{E}(\bar{X}) = \mathbb{E}(X_1) = \lambda$ and $\mathbb{V}ar(\bar{X}) = \mathbb{V}ar(X_1)/n = \lambda/n \to \infty$ as $n$ increases. $\square$

**Example (uniform distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} Unif(0, \theta)$. Consider first the MM estimator $2\bar{X}$. Its mean is $\mathbb{E}(2\bar{X}) = 2\mathbb{E}(X_1) = (2)(\theta/2) = \theta$, so this estimator is unbiased. Further, this estimator is consistent since

$$\mathbb{V}ar(2\bar{X}) = 4 \cdot \frac{\mathbb{V}ar(X_1)}{n} = \frac{4}{n} \cdot \frac{\theta^2}{12} = \frac{\theta^2}{3n} \to 0, \quad \text{as } n \to \infty.$$

Next, consider the MLE estimator $X_{(n)}$. We will show that it is a biased estimator of $\theta$ and modify it to give an unbiased estimator. We start by finding the cdf of $X_{(n)}$. The cdf of a $Unif(0,\theta)$ random variable is $F(x) = x/\theta,\ 0 \le x \le \theta$, and therefore, the cdf of the maximum is $F_{X_{(n)}}(x;\theta) = F^n(x) = \dfrac{x^n}{\theta^n},\ \ 0 \le x \le \theta$. From here, the density of $X_{(n)}$ is $f_{X_{(n)}}(x;\theta) = F'_{X_{(n)}}(x;\theta) = nx^{n-1}/\theta^n,\ \ 0 \le x \le \theta$. And thus the expected value is derived as

$$\mathbb{E}(X_{(n)}) = \int_0^\theta x\, n\, \frac{x^{n-1}}{\theta^n}\, dx = \frac{n}{n+1}\,\theta = \big(1 - \frac{1}{n+1}\big)\theta < \theta.$$

We can see that $X_{(n)}$ is a biased estimator of $\theta$, and, in fact, it underestimates $\theta$ by $1/(n+1)$th of $\theta$, on average. An unbiased estimator of $\theta$ based on the maximum value is $\dfrac{n+1}{n}\, X_{(n)}$.

**Note:** If $\widehat{\theta}$ is biased and $\mathbb{E}(\widehat{\theta}) = c\theta$, then $\widehat{\theta}/c$ is an unbiased estimator of $\theta$ since $\mathbb{E}(\widehat{\theta}/c) = (1/c)\,\mathbb{E}(\widehat{\theta}) = (1/c)(c\theta) = \theta$.

Finally, because $X_{(n)}$ is biased, it is not consistent. However, the unbiased estimator $\dfrac{n+1}{n}\, X_{(n)}$ is a consistent estimator of $\theta$ as shown via some algebraic manipulations. We write $\mathbb{V}ar\Big(\dfrac{n+1}{n}\, X_{(n)}\Big) = \Big(\dfrac{n+1}{n}\Big)^2 \mathbb{V}ar(X_{(n)})$. Next, $\mathbb{V}ar(X_{(n)}) = \mathbb{E}(X_{(n)}^2) - \Big(\mathbb{E}(X_{(n)})\Big)^2 = \displaystyle\int_0^\theta x^2\, \frac{nx^{n-1}}{\theta^n}\, dx - \Big(\frac{n}{n+1}\theta\Big)^2 = \Big(\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\Big)\theta^2 = n\theta^2\Big(\dfrac{n^2+2n+1-n^2-2n}{(n+2)(n+1)^2}\Big) = \dfrac{n\theta^2}{(n+2)(n+1)^2}$. Thus, $\mathbb{V}ar\Big(\dfrac{n+1}{n}\, X_{(n)}\Big) = \Big(\dfrac{n+1}{n}\Big)^2 \mathbb{V}ar(X_{(n)}) = \dfrac{\theta^2}{n(n+2)} \to 0$ as $n \to \infty$, and so the estimator is consistent. $\square$

**Example (exponential distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} Exp$ with mean $\beta$. The MLE and MM estimators of $\beta$ are $\bar{X}$, and as noted earlier, it is an unbiased and consistent estimator. $\square$

**Example (normal distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. The MLE and MM estimator of $\mu$, $\bar{X}$, is unbiased and consistent. The MLE and MM estimator of $\sigma^2$ is $\widehat{\sigma}^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^n (X_i - \bar{X})^2$. We will show that it is biased. We write

$$\mathbb{E}(\widehat{\sigma}^2) = \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\Big] = \frac{1}{n}\,\mathbb{E}\big(\sum_{i=1}^n X_i^2 - n\,\bar{X}^2\big)$$

$$= \mathbb{E}(X_1^2) - \mathbb{E}(\bar{X}^2) = \mathbb{V}ar(X_1) + (\mathbb{E}(X_1))^2 - \big[\mathbb{V}ar(\bar{X}) + (\mathbb{E}(\bar{X}))^2\big]$$

$$= \sigma^2 + \mu^2 - \big(\frac{\sigma^2}{n} + \mu^2\big) = \frac{n-1}{n}\,\sigma^2.$$

Hence, $s^2 = \dfrac{n}{n-1}\widehat{\sigma}^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is an unbiased estimator of $\sigma^2$. Its variance is equal to $\dfrac{2\sigma^4}{n-1}$. The easiest way to see that is by using the fact that $\dfrac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$. The variance of a random variable with a $\chi^2(k)$ distribution is $2k$. Therefore, $\mathbb{V}ar(s^2) = \dfrac{\sigma^4}{(n-1)^2}\mathbb{V}ar\Big(\dfrac{(n-1)s^2}{\sigma^2}\Big) = \dfrac{\sigma^4}{(n-1)^2}\cdot 2(n-1) = \dfrac{2\sigma^4}{n-1} \to 0$, as $n \to \infty$. Hence, $s^2$ is a consistent estimator of $\sigma^2$. $\square$

## PROPERTY OF ESTIMATORS: SUFFICIENCY

**Definition.** Any function of observations $x_1, \ldots, x_n$ is called **statistic**.

Note that a parameter estimator $\widehat{\theta}$ is a statistic since it depends on observations. We write $\widehat{\theta} = \widehat{\theta}(x_1, \ldots, x_n)$.

Here is a formal definition of a sufficient statistic.

**Definition.** Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x|\theta)$. A statistic $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ is called a **sufficient statistic** for $\theta$ if the conditional distribution of $X_1, \ldots, X_n$ given $\widehat{\theta}$ doesn't depend on $\theta$.

On an intuitive level, a statistic is sufficient for $\theta$ if it alone can be used to estimate $\theta$. No additional information about the sample is needed.

**Example.** Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$. Note that $\bar{X} \sim N(\mu, 1/n)$. We know that $\bar{X}$ is an estimator of $\mu$, so $\bar{X}$ (or, equivalently, $(\sum_{i=1}^n X_i,\ n)$) is a sufficient statistic for $\mu$. We can see that it satisfies the formal definition. The conditional joint density of $X_1, \ldots, X_n$ given $\bar{X}$ can be found as follows.

$$f_{X_1,\ldots,X_n|\bar{X}}(x_1, \ldots, x_n \,|\, \bar{x}) = \frac{f_{X_1}(x_1)\cdots f_{X_n}(x_n)}{f_{\bar{X}}(\bar{x})} \quad (\text{where } x_1 + \cdots + x_n = n\bar{x})$$

$$= \frac{(2\pi)^{-n/2}}{(2\pi/n)^{-1/2}} \exp\Big\{ -\frac{1}{2}\Big((x_1-\mu)^2 + (x_n-\mu)^2 - n(\bar{x}-\mu)^2\Big)\Big\}$$

$$= \sqrt{n}\,(2\pi)^{-(n-1)/2} \exp\Big\{ -\frac{1}{2}\Big(x_1^2 + \cdots + x_n^2 - 2n\bar{x}\,\mu + n\mu^2 - n\bar{x}^2 + 2n\bar{x}\,\mu - n\mu^2\Big)\Big\}$$

$$= \sqrt{n}\,(2\pi)^{-(n-1)/2} \exp\Big\{ -\frac{1}{2}\Big(x_1^2 + \cdots + x_n^2 - n\bar{x}^2\Big)\Big\},$$

which doesn't depend on $\mu$. $\square$

In practice, to find a sufficient statistic for a parameter, the following theorem is utilized.

**Factorization Theorem.** Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\, \theta)$. Then $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ is a sufficient statistic for $\theta$ if and only if there exist functions $g$ and $h$ such that

$$\prod_{i=1}^{n} f(X_i \,|\, \theta) = g(X_1, \ldots, X_n)\, h(\widehat{\theta}, \theta).$$

**Proposition.** Any function of a sufficient statistic is sufficient.

**Example.** By this proposition, if, for instance, $\sum_{i=1}^{n} X_i$ is sufficient, then $\bar{X}$ is also sufficient. □

**Example (Bernoulli distribution).** Suppose $X_1, \ldots, X_n \overset{iid}{\sim} Ber(p)$. To find a sufficient statistic for $p$, we apply the Factorization theorem. We write

$$\prod_{i=1}^{n} f(X_i \,|\, p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\sum X_i}(1-p)^{n-\sum X_i}.$$

Now, since $\sum X_i$ cannot be factored out (that is, separated from $p$ in a multiplicative fashion), it has to be a sufficient statistic. We have $\widehat{p} = \sum X_i$, $h(\widehat{p}, p) = p^{\widehat{p}}(1-p)^{n-\widehat{p}}$ and $g(X_1, \ldots, X_n) \equiv 1$. Note that here we assume that $n$ is a known constant. Also, by the above proposition, since $\sum X_i$ is sufficient for $p$, $\bar{X}$ is also sufficient. □

**Example (geometric distribution).** Suppose $X_1, \ldots, X_n \overset{iid}{\sim} Geom(p)$. Using the Factorization theorem, we write

$$\prod_{i=1}^{n} f(X_i \,|\, p) = \prod_{i=1}^{n} p(1-p)^{X_i - 1} = p^n (1-p)^{\sum X_i - n}.$$

Here $\widehat{p} = \sum X_i$ is a sufficient statistic, $h(\widehat{p}, p) = p^n(1-p)^{\widehat{p}-n}$ and $g(X_1, \ldots, X_n) \equiv 1$. Consequently, $\bar{X}$ is also sufficient. □

**Example (Poisson distribution).** Suppose $X_1, \ldots, X_n \overset{iid}{\sim} Poi(\lambda)$. Applying the Factorization theorem, we get

$$\prod_{i=1}^{n} f(X_i \,|\, \lambda) = \prod_{i=1}^{n} \frac{\lambda^{X_i}}{X_i!}\, e^{-\lambda} = \Big(\prod_{i=1}^{n} X_i!\Big)^{-1} \lambda^{\sum X_i}\, e^{-n\lambda}.$$

Denoting by $\widehat{\lambda} = \sum X_i$ a sufficient statistic for $\lambda$, we have $h(\widehat{\lambda}, \lambda) = \lambda^{\sum X_i}\, e^{-n\lambda}$ and $g(X_1, \ldots, X_n) = \big(\prod_{i=1}^{n} X_i!\big)^{-1}$. Consequently, $\bar{X}$ is also sufficient. □

**Example (uniform distribution).** Suppose $X_1, \ldots, X_n \overset{iid}{\sim} Unif(0, \theta)$. By the Factorization theorem,

$$\prod_{i=1}^{n} f(X_i \,|\, \theta) = \prod_{i=1}^{n} \frac{1}{\theta}\, \mathbb{I}(0 \le X_i \le \theta) = \mathbb{I}(0 \le X_{(1)})\, \frac{1}{\theta^n} \mathbb{I}(X_{(n)} \le \theta).$$

We see that $\widehat{\theta} = X_{(n)}$ is sufficient for $\theta$. The functions are $h(\widehat{\theta}, \theta) = \dfrac{1}{\theta^n} \mathbb{I}(\widehat{\theta} \le \theta)$ and $g(X_1, \ldots, X_n) = \mathbb{I}(0 \le X_{(1)})$. $\square$

**Example (exponential distribution).** Let $X_1, \ldots, X_n$ be iid exponential random variables with mean $\beta$. We use the Factorization Theorem and write

$$\prod_{i=1}^{n} f(X_i \,|\, \beta) = \prod_{i=1}^{n} \frac{1}{\beta}\, e^{-X_i/\beta} = \frac{1}{\beta^n}\, \exp\{-\sum X_i/\beta\}.$$

Now, we identify $\widehat{\beta} = \sum X_i$ as a sufficient statistic for $\beta$ that gives $h(\widehat{\beta}, \beta) = \frac{1}{\beta^n}\, \exp\{-\widehat{\beta}/\beta\}$ and $g(X_1, \ldots, X_n) \equiv 1$. Hence, $\bar{X}$ is also sufficient. $\square$

**Example (normal distribution).** Assume $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Applying the Factorization Theorem, we write

$$\prod_{i=1}^{n} f(X_i \,|\, \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{ - \frac{(X_i - \mu)^2}{2\sigma^2} \Big\}$$

$$= (2\pi)^{-n/2}\, (\sigma^2)^{-n/2}\, \exp\Big\{ - \sum X_i^2/(2\sigma^2) + (\mu/\sigma^2) \sum X_i - n\mu^2/(2\sigma^2) \Big\}.$$

It is clear that $\widehat{\mu} = \sum X_i$ is sufficient for $\mu$ and $\widehat{\sigma}^2 = (\sum X_i, \sum X_i^2)$ is sufficient for $\sigma^2$. The function $g$ in this case is $g(X_1, \ldots, X_n) = (2\pi)^{-n/2}$. As a result, $\bar{X}$ is sufficient for $\mu$, and

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

is sufficient for $\sigma^2$. $\square$

Why do we need sufficient statistics? In a nutshell, unbiased sufficient statistics have the smallest variance among all unbiased statistics and thus are the best possible. Before formulating this statement in the form of a theorem, let's consider an example of a shifted exponential distribution.

**Example (shifted exponential distribution).** Consider $X_1, \ldots, X_n \overset{iid}{\sim} f(x \,|\, \theta) = e^{-(x-\theta)}, \quad x > \theta$. First, we find the method of moments estimator and study its properties. We have

$$\mathbb{E}(X_1) = \int_\theta^\infty x\, e^{-(x-\theta)}\, dx = \{y = x-\theta\} = \int_0^\infty (y+\theta)\, e^{-y}\, dy = \int_0^\infty y\, e^{-y}\, dy + \theta \int_0^\infty e^{-y}\, dy \;=\; 1+\theta,$$

hence $1+\widehat{\theta}_{MM} = \bar{X}$ and $\widehat{\theta}_{MM} = \bar{X} - 1$. It is an unbiased estimator of $\theta$ since $\mathbb{E}(\widehat{\theta}_{MM}) = \mathbb{E}(\bar{X}-1) = \mathbb{E}(X_1) - 1 = 1 + \theta - 1 = \theta$, Its variance is equal to

$$\mathbb{V}ar(\widehat{\theta}_{MM}) = \mathbb{V}ar(\bar{X} - 1) = \mathbb{V}ar(\bar{X}) = \frac{\mathbb{V}ar(X_1)}{n}$$
$$= \frac{1}{n}\int_\theta^\infty (x - \theta - 1)^2\, e^{-(x-\theta)}\, dx = \{y = x - \theta\} = \frac{1}{n}\int_0^\infty (y-1)^2\, e^{-y}\, dy \;=\; \frac{1}{n}.$$

Consider now the maximum likelihood estimator. We have shown earlier that $\widehat{\theta}_{MLE} = X_{(1)}$. Its cdf is

$$F_{X_{(1)}}(x) = 1 - \mathbb{P}(X_{(1)} \geq x) = 1 - \mathbb{P}(X_1 \geq x) \cdot \cdots \cdot \mathbb{P}(X_n \geq x) = 1 - e^{-n(x-\theta)}, \;\; x > \theta.$$

The pdf is $f_{X_{(1)}}(x) = n\, e^{-n(x-\theta)},\; x > \theta$. The mean is

$$\mathbb{E}(X_{(1)}) = \int_\theta^\infty x\, n\, e^{-n(x-\theta)}\, dx = \{y = n(x-\theta), dy = ndx\} = \int_0^\infty (y + n\theta)\, e^{-y}\, \frac{1}{n}\, dy$$
$$= \frac{1}{n}\int_0^\infty y\, e^{-y}\, dy \;+\; \theta \int_0^\infty e^{-y}\, dy \;=\; \frac{1}{n} \;+\; \theta.$$

The variance is calculated as

$$\mathbb{V}ar(X_{(1)}) = \int_\theta^\infty (x - \frac{1}{n} - \theta)^2\, n\, e^{-n(x-\theta)}\, dx = \{y = n(x-\theta), dy = ndx\} = \int_0^\infty \frac{1}{n^2}\, (y-1)^2\, e^{-y}\, dy \;=\; \frac{1}{n^2}.$$

An unbiased estimator that is based on $\widehat{\theta}_{MLE} = X_{(1)}$ is $\widehat{\theta} = X_{(1)} - \frac{1}{n}$. It is unbiased and its variance is $\mathbb{V}ar(\widehat{\theta}) = \mathbb{V}ar(X_{(1)}) = \frac{1}{n^2}$. Note that $\widehat{\theta}_{MM} = \bar{X} - 1$ and $\widehat{\theta} = X_{(1)} - \frac{1}{n}$ and both unbiased but $\mathbb{V}ar(\widehat{\theta}) = \dfrac{1}{n^2} < \dfrac{1}{n} = \mathbb{V}ar(\widehat{\theta}_{MM})$ for all $n > 1$. It means that for every fixed $n > 1$, $\widehat{\theta}$ performs better than $\widehat{\theta}_{MM}$. Why is it so? Because $\widehat{\theta}$ is based on a sufficient statistic $X_{(1)}$. To see this, we apply the Factorization Theorem. We have

$$\prod_{i=1}^n f(X_i \,|\, \theta) \;=\; \prod_{i=1}^n e^{-(X_i - \theta)}\, \mathbb{I}(X_i > \theta) \;=\; \exp\{-\sum X_i + n\theta\}\, \mathbb{I}(X_{(1)} > \theta).$$

We see that $X_{(1)}$ has to be a sufficient statistic. The functions are $g(X_1, \ldots, X_n) = \exp\{-\sum X_i\}$ and $h(X_{(1)}, \theta) = \exp\{n\theta\}\, \mathbb{I}(X_{(1)} > \theta)$. $\square$

Further, we state a theorem that explains not only what estimators perform better than others but also how to find the best-performing estimators.

# THE RAO-BLACKWELL THEOREM

**Theorem (Rao-Blackwell (R-B) Theorem).** Let $X_1, \ldots, X_n$ be iid with pmf or pdf $f(x;\theta)$, and let $u = u(X_1, \ldots, X_n)$ be a sufficient statistic for $\theta$. Suppose $\widehat{\theta}$ is an unbiased estimator of $\theta$. Consider a new estimator of $\theta$ defined as the conditional expected value

$$\mathbb{E}(\widehat{\theta}\,|\,u) = \begin{cases} \sum_{\widehat{\theta}=x} x\,\mathbb{P}(\widehat{\theta}=x\,|\,u), & \text{in discrete case} \\ \int_{-\infty}^{\infty} x\, f_{\widehat{\theta}}(x\,|\,u)\,dx, & \text{in continuous case.} \end{cases}$$

This estimator has the following properties: (i) it depends on the sufficient statistic $u$ (by the definition of conditional expectation); (ii) it is unbiased since $\mathbb{E}\big(\mathbb{E}(\widehat{\theta}\,|\,u)\big) = \mathbb{E}(\widehat{\theta}) = \theta$ (by the double-expectation formula); and (iii) its variance is smaller than the variance of $\widehat{\theta}$, that is, $\mathbb{V}ar\big(\mathbb{E}(\widehat{\theta}\,|\,u)\big) \leq \mathbb{V}ar(\widehat{\theta})$ (that's the main assertion of this theorem which we will not prove here).

**Note.** Calyampudi Radhakrishna (CR) Rao (b. 1920) is a celebrated Indian-American mathematician and statistician (yes, aged 102), and David Blackwell (1919-2010) was a renown African-American mathematician and statistician.

**Definition.** An unbiased estimator with the smallest variance is called the **uniform minimum variance unbiased estimator** (UMVUE) or the **best estimator**.

**Remark.** If an unbiased estimator of $\theta$ already depends on a sufficient statistic $u$, then it the UMVUE and its performance cannot be improved. This is so because $\mathbb{E}(\widehat{\theta}(u)\,|\,u) = \widehat{\theta}(u)$.

**Remark.** Given a sufficient statistic $u$ and any unbiased estimator $\widehat{\theta}$ of $\theta$, we can produce the UMVUE of $\theta$ by computing the conditional expectation $\mathbb{E}(\widehat{\theta}\,|\,u)$. Generally speaking, computing this conditional expectation is a formidable task, but in some special cases it is relatively easy and explicit results exist. We will consider some examples later.

**Remark.** The UMVUE is necessarily unique. This statement needs a proof which is omitted here.

**Example (Bernoulli distribution).** Let $X_1, \ldots, X_n$ be iid $Ber(p)$ random variables. We know that $\widehat{p}_{MLE} = \widehat{p}_{MM} = \bar{X}$ and $\bar{X}$ is an unbiased estimator and a sufficient statistic for $p$. Therefore, by the R-B theorem, $\bar{X}$ is the UMVUE of $p$. $\square$

**Example (geometric distribution).** Suppose $X_1, \ldots, X_n \overset{iid}{\sim} Geom(p)$. We have proven earlier that $\widehat{p}_{MLE} = \widehat{p}_{MM} = 1/\bar{X}$ is a biased estimator of $p$. It means that it cannot be the best estimator of $p$. $\square$

**Example (Poisson distribution).** Let $X_1, \ldots, X_n$ be iid $Poi(\lambda)$ random variables. We have shown that $\widehat{\lambda}_{MLE} = \widehat{\lambda}_{MM} = \bar{X}$ which is unbiased and sufficient and hence, it is the UMVUE of $\lambda$. $\square$

**Example (uniform distribution).** Consider $X_1, \ldots, X_n \overset{iid}{\sim} Unif(0, \theta)$. We know that $\widehat{\theta}_{MLE} = X_{(n)}$ is a biased estimator of $\theta$, but $\dfrac{n+1}{n} X_{(n)}$ is unbiased. Also, $X_{(n)}$ is a sufficient statistic. Therefore, by the R-B theorem, $\dfrac{n+1}{n} X_{(n)}$ is the UMVUE estimator of $\theta$.

Consider now the method of moments estimator $\widehat{\theta}_{MM} = 2\bar{X}$. We know that it is an unbiased estimator of $\theta$, but it is not based on a sufficient statistic $X_{(n)}$ and thus, by the R-B theorem, it can be improved upon by computing the conditional expectation $\mathbb{E}\big(2\bar{X} \,|\, X_{(n)}\big)$, which will be the UMVUE of $\theta$. Note that we know that the UMVUE is unique, and $\dfrac{n+1}{n} X_{(n)}$ is the UMVUE of $\theta$. Hence, we know that $\mathbb{E}\big(2\bar{X} \,|\, X_{(n)}\big)$ must equal to $\dfrac{n+1}{n} X_{(n)}$. To verify that, we consider two cases: (i) $X_{(n)} = X_1$ which happens with probability $1/n$, and (i) $X_1 < X_{(n)}$ which has probability $1 - 1/n$, and in this case, $X_1 \sim Unif(0,\, X_{(n)})$. We write

$$\mathbb{E}\big(2\bar{X} \,|\, X_{(n)}\big) \;=\; 2\mathbb{E}\big(X_1 \,|\, X_{(n)}\big) = 2\Big(X_{(n)}\frac{1}{n} \;+\; \frac{X_{(n)}}{2}\,\big(1 - \frac{1}{n}\big)\Big)$$

$$= \frac{2X_{(n)}}{n} \;+\; X_{(n)} \;-\; \frac{X_{(n)}}{n} \;==\; \frac{X_{(n)}}{n} \;+\; X_{(n)} \;=\; \frac{n+1}{n} X_{(n)}. \quad \square$$

**Example (exponential distribution).** Consider $X_1, \ldots, X_n$ that are independent and exponentially distributed with mean $\beta$. We have shown that $\widehat{\beta}_{MLE} = \widehat{\beta}_{MM} = \bar{X}$ is an unbiased estimator of $\beta$ and $\bar{X}$ is sufficient, therefore, it is the UMVUE estimator of $\beta$. $\square$

**Example (normal distribution).** Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ random variables. We know that $\widehat{\mu}_{MLE} = \widehat{\mu}_{MM} = \bar{X}$ is unbiased and sufficient and therefore, is the best estimator (UMVUE) of $\mu$.

We also know that $\widehat{\sigma}^2_{MLE} = \widehat{\sigma}^2_{MM} = \dfrac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2$ is biased but $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$ is an unbiased estimator of $\sigma^2$. It is based on a sufficient statistic $(\sum X_i, \sum X_i^2)$, and therefore, $s^2$ is the UMVUE of $\sigma^2$.

Further, recall that we have shown earlier that the variance of $s^2$ is equal to $\dfrac{2\sigma^4}{n-1}$. The proof was based on the fact that $\dfrac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$ which variance is $2(n-1)$. By the same token, we can find the variance of $\widehat{\sigma}^2_{MLE}$. It is computed as

$$\mathbb{V}ar(\widehat{\sigma}^2_{MLE}) = \mathbb{V}ar\big(\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\big) \;=\; \frac{1}{n^2}\,\mathbb{V}ar\big(\sum_{i=1}^n (X_i - \bar{X})^2\big)$$

$$= \;\frac{\sigma^4}{n^2}\,\mathbb{V}ar\Big(\frac{(n-1)s^2}{\sigma^2}\Big) \;=\; \frac{\sigma^4}{n^2}\cdot 2(n-1) \;=\; 2\sigma^4\frac{n-1}{n^2} \;<\; 2\sigma^4\frac{1}{n-1} \;=\; \mathbb{V}ar(s^2) \;\text{ for all }\; n>1.$$

How can it happen that $s^2$ is the UMVUE and thus has the smallest possible variance, but we produced an estimator with a smaller variance? The answer is that the estimator with the smaller variance is a biased estimator of $\sigma^2$. This illustrates that it is possible to come up with an estimator with a smaller variance than the UMVUE but at the expense of unbiasedness. $\square$

Next, we consider a few examples, where the explicit form of the UMVUE is hard to guess but it can be found by computing the conditional expectation in the R-B Theorem.

**Example.** Let $X_1, \ldots, X_n$ be iid $Poi(\lambda)$ random variables. Suppose we would like to find the UMVUE of $p(0) = \mathbb{P}(X_1 = 0) = e^{-\lambda}$. A natural candidate to try is the MLE $e^{-\widehat{\lambda}} = e^{-\bar{X}}$ but it is a biased estimator of $p(0)$. To convince ourselves, we compute its mean. Recall that the moment generating function of $X \sim Poi(\lambda)$ is

$$M_X(t) \stackrel{def}{=} \mathbb{E}\big(e^{tX}\big) \;=\; \sum_{x=0}^{\infty} \frac{e^{tx}\lambda^x}{x!}\,e^{-\lambda} \;=\; e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda\,e^t)^x}{x!} \;=\; e^{-\lambda}\,\exp\{\lambda\,e^t\} = \exp\{\lambda(e^t-1)\}.$$

Thus,

$$\mathbb{E}\big(e^{-\bar{X}}\big) = \mathbb{E}\big(e^{-\frac{1}{n}(X_1+\cdots X_n)}\big) = \{\text{by independence}\} = \mathbb{E}\big(e^{-\frac{1}{n}X_1}\big)\cdot\cdots\cdot\mathbb{E}\big(e^{-\frac{1}{n}X_n}\big)$$

$$= \;\Big(\exp\{\lambda\big(e^{-1/n}-1\big)\}\Big)^n \;=\; \exp\{n\lambda\big(e^{-1/n}-1\big)\} \;\neq\; e^{-\lambda}.$$

Note that $e^{-\bar{X}}$ is **asymptotically unbiased** since for large $n$, $\mathbb{E}\big(e^{-\bar{X}}\big) \;=\; \exp\{n\lambda\big(e^{-1/n}-1\big)\} \;\approx$ $\exp\big\{n\lambda\big(1-\frac{1}{n}-1\big)\big\} \;=\; e^{-\lambda}$. However, the UMVUE has to be unbiased for any fixed $n$.

Further, to find the UMVUE of $e^{-\lambda}$ we need to take any unbiased estimator and compute its expected value conditioned on the value of a sufficient statistic $\bar{X}$. Since we want to estimate a theoretic probability of zero, an estimator that always works as its unbiased estimator is the empirical estimator $\widehat{P}(X_1=0) = \dfrac{\#\;\text{of zeros in the sample}}{n} = \dfrac{\sum_{i=1}^n \mathbb{I}(X_i=0)}{n}$. Indeed, its mean is

$$\mathbb{E}\Big(\frac{\sum_{i=1}^n \mathbb{I}(X_i=0)}{n}\Big) \;=\; \mathbb{E}\big(\mathbb{I}(X_1=0)\big) = (1)\mathbb{P}(X_1=0) + (0)\mathbb{P}(X_1\neq 0) = \mathbb{P}(X_1=0).$$

To produce the UMVUE of $\mathbb{P}(X_1=0)$, we assume that $\sum_{i=1}^n X_i = x$ and compute the conditional expectation $\mathbb{E}\Big(\dfrac{\sum_{i=1}^n \mathbb{I}(X_i=0)}{n} \;\Big|\; \sum_{i=1}^n X_i = x\Big)$. The final answer will depend on $x$ which we will

replace with $\sum_{i=1}^n X_i$ or better $n\bar{X}$. We write

$$\mathbb{E}\Big(\frac{\sum_{i=1}^n \mathbb{I}(X_i=0)}{n} \,\Big|\, \sum_{i=1}^n X_i = x\Big) \;=\; \mathbb{E}\big(\mathbb{I}(X_1=0) \,|\, \sum_{i-1}^n X_i = x\big)$$

$$= (1)\mathbb{P}(X_1=0 \,|\, \sum_{i=1}^n X_i = x) + (0)\mathbb{P}(X_1\neq 0 \,|\, \sum_{i=1}^n X_i = x) \;=\; \mathbb{P}(X_1=0 \,|\, \sum_{i=1}^n X_i = x)$$

$$= \frac{\mathbb{P}(X_1=0,\, \sum_{i=1}^n X_i = x)}{\mathbb{P}(\sum_{i=1}^n X_i = x)} \;=\; \frac{\mathbb{P}(X_1=0,\, \sum_{i=2}^n X_i = x)}{\mathbb{P}(\sum_{i=1}^n X_i = x)}$$

$$= \{\text{independence}\} \;=\; \frac{\mathbb{P}(X_1=0)\,\mathbb{P}(\sum_{i=2}^n X_i = x)}{\mathbb{P}(\sum_{i=1}^n X_i = x)}$$

$$= \Big\{\sum_{i=2}^n X_i \sim Poi\big((n-1)\lambda\big),\, \sum_{i=1}^n X_i \sim Poi(n\lambda)\Big\} \;=\; \frac{e^{-\lambda}\cdot \frac{((n-1)\lambda)^x}{x!}\, e^{-(n-1)\lambda}}{\frac{(n\lambda)^x}{x!}\, e^{-n\lambda}}$$

$$= \Big(1-\frac{1}{n}\Big)^x \;=\; \Big(1-\frac{1}{n}\Big)^{n\bar{X}}.$$

We procured the UMVUE of $e^{-\lambda}$. It is $\Big(1-\frac{1}{n}\Big)^{n\bar{X}}$. Note that as $n$ increases, it tends to $e^{-\bar{X}}$. Thus, in the limit, it coincides with the MLE estimator. $\square$

**Example.** Let $X_1,\ldots,X_n \overset{iid}{\sim} Ber(p)$. Suppose we need to find the UMVUE of $p^2$. We can try an MLE-based estimator $\widehat{p^2} = \bar{X}^2$. Its mean is $\mathbb{E}(\bar{X}^2) = \mathbb{V}ar(\bar{X}) + (\mathbb{E}\bar{X})^2 \;=\; \dfrac{\mathbb{V}ar(X_1)}{n} + (\mathbb{E}X_1)^2 \;=\; \dfrac{p(1-p)}{n} + p^2 \neq p^2$, and thus, it is not an unbiased estimator.
We now resort to the R-B theorem. As an unbiased estimator of $p^2$ we will take $X_1\cdot X_2$. Indeed, $\mathbb{E}(X_1\cdot X_2) = \{\textit{independence}\} = \mathbb{E}X_1\cdot\mathbb{E}X_2 = p\cdot p = p^2$. Conditioning on a sufficient statistic $\sum X_i = x$, we obtain

$$\mathbb{E}\big(X_1\cdot X_2 \,|\, \sum_{i=1}^n X_i = x\big) \;=\; \mathbb{P}\big(X_1=1, X_2=1 \,|\, \sum_{i=1}^n X_i = x\big)$$

$$= \frac{\mathbb{P}\big(X_1=1, X_2=1, \sum_{i=3}^n X_i = x-2\big)}{\mathbb{P}\big(\sum_{i=1}^n X_i = x\big)}$$

$$= \{\textit{independence}\} = \frac{\mathbb{P}(X_1=1)\,\mathbb{P}(X_2=1)\,\mathbb{P}\big(\sum_{i=3}^n X_i = x-2\big)}{\mathbb{P}\big(\sum_{i=1}^n X_i = x\big)}$$

$$= \{X_1\sim Ber(p),\; X_2\sim Ber(p),\; \sum_{i=1}^n X_i \sim Bi(n,p),\; \sum_{i=3}^n X_i \sim Bi(n-2,p)\}$$

$$= \frac{p \cdot p \cdot \binom{n-2}{x-2} p^{x-2}(1-p)^{(n-2)-(x-2)}}{\binom{n}{x} p^x (1-p)^{n-x}} = \frac{\binom{n-2}{x-2}}{\binom{n}{x}}$$

$$= \frac{(n-2)!}{(x-2)!(n-x)!} \cdot \frac{x!(n-x)!}{n!} = \frac{x(x-1)}{n(n-1)} = \frac{n\bar{X}(n\bar{X}-1)}{n(n-1)} = \frac{\bar{X}(n\bar{X}-1)}{n-1}$$

is the UMVUE of $p^2$. Note that as $n$ increases, this estimator goes to the MLE $\bar{X}^2$. $\square$

**Example.** Let $X_1, \dots, X_n$ be iid exponential random variables with mean $\beta$. Suppose we need to find $\mathbb{P}(X_1 \le 2) = 1 - e^{-2/\beta}$. The MLE estimator $1 - e^{2/\bar{X}}$ is not unbiased. We turn to the R-B theorem. An unbiased estimator of $\mathbb{P}(X_1 \le 2)$ is the empirical estimator $\sum_{i=1}^{n} \mathbb{I}(X_i \le 2)/n$. Conditioning on a sufficient statistic $\sum X_i = x$, we compute

$$\mathbb{E}\Big( \frac{\sum_{i=1}^n \mathbb{I}(X_i \le 2)}{n} \,\Big|\, \sum_{i=1}^n X_i = x \Big) = \mathbb{E}\big( \mathbb{I}(X_1 \le 2) \,\big|\, \sum_{i=1}^n X_i = x \big)$$

$$= \int_0^2 f_{X_1 | \sum_{i=1}^n X_i}(u|x)\, du = \int_0^2 \frac{f_{X_1, \sum_{i=1}^n X_i}(u, x)}{f_{\sum_{i=1}^n X_i}(x)}\, du$$

$$= \int_0^2 \frac{f_{X_1, \sum_{i=2}^n X_i}(u, x-u)}{f_{\sum_{i=1}^n X_i}(x)}\, du = \{independence\} = \int_0^2 \frac{f_{X_1}(u)\, f_{\sum_{i=2}^n X_i}(x-u)}{f_{\sum_{i=1}^n X_i}(x)}\, du$$

$$= \Big\{ \sum_{i=1}^n X_i \sim Gamma(n, \beta),\ \sum_{i=2}^n X_i \sim Gamma(n-1, \beta) \Big\}$$

$$= \int_0^2 \frac{\frac{1}{\beta} e^{-u/\beta} \frac{(x-u)^{n-2}}{(n-2)!\beta^{n-1}} e^{-(x-u)/\beta}}{\frac{x^{n-1}}{(n-1)!\beta^n} e^{-x/\beta}}\, du = \frac{n-1}{x^{n-1}} \int_0^2 (x-u)^{n-2}\, du$$

$$= \{y = u/x,\, dy = du/x\} = (n-1) \int_0^{2/x} (1-y)^{n-2}\, dy = -(1-y)^{n-1} \Big|_0^{2/x} = 1 - (1 - 2/x)^{n-1}.$$

Therefore, the UMVUE of $\mathbb{P}(X_1 \le 2)$ is $1 - \Big(1 - \frac{2}{n\bar{X}}\Big)^{n-1}$. Note that as $n$ goes to infinity, this estimator approaches the MLE $1 - e^{-2/\bar{X}}$. $\square$

## INTERVAL ESTIMATOR

The estimators that we have been studying so far are called **point estimators** because they estimate an unknown parameter by a single value computed from the given sample. The major drawbacks of point estimators are that they depend on data (for different samples point estimators are different), and that the true parameter value is not equal to the estimated value. It is better to compute interval estimators.

**Definition.** Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x|\theta)$, and let $\widehat{\theta}$ be the point estimator of $\theta$. An **interval estimator** of $\theta$ has the form $\widehat{\theta} \pm margin\ of\ error$.

**Note.** The margin of error is typically denoted by $m$. The endpoints of an interval estimator are $\widehat{\theta} - m$ and $\widehat{\theta} + m$. The length of an interval estimator is $2m$.

**Note.** A theoretical construct $\widehat{\theta}$ is referred to as **estimator**, whereas an empirically computed value is termed **estimate**.

**Properties of Interval Estimators.** (1) For some samples, interval estimates will cover the true population parameter, but for some samples, the interval estimates will be off. We want to limit these mistakes to some small pre-determined percent of the samples. For example, we might want 95% of the interval estimates to cover the true parameter, and, respectively, 5% to be off. Or we might want to increase the percentage of correct coverage to 99% and erroneous interval estimates not to exceed 1%. What should be true about interval estimators is that intervals with larger percent coverage (respectively, smaller percent of error) should be wider. For example, intervals with 99% true coverage should be wider than those with 95% true coverage. Indeed, if we take 5% of erroneous intervals and compute a wider interval, then some of them will cover the true parameter, leaving only 1% non-coverage.

(2) If we increase a sample size, the margin of error of an interval estimator should become smaller (and thus, the interval should become narrower). If we sample the entire population, then we would know the exact value of the parameter and so the interval estimator would collapse to a point.

**Definition.** A $100(1-\alpha)\%$ **confidence interval** (CI) for $\theta$ is of the form $\widehat{\theta} \pm m$, where $m$ is the margin of error. For $100(1-\alpha)\%$ of samples, the CI covers the true parameter $\theta$, and for $100\alpha\%$ of samples, it doesn't cover. The quantity $100(1-\alpha)\%$ is called **confidence level**.

**Note.** In view of the properties of interval estimators, margin of error of a CI should increase as confidence level increases, and decrease as sample size increases.

**Note.** Since a population parameter has a fixed value but endpoints of a CI are random (depend on sample), it is correct to say that "CI covers the true parameter" and it is incorrect to say that

"parameter lies within the CI".

**Note.** CIs that are mostly commonly used in practice are 90%, 95%, and 99% (with 95% CI being the most widespread). The corresponding values of $\alpha$ are 0.1 for 90% CI, 0.05 for 95% CI, and 0.01 for 99% CI. Sometimes, especially in medical field, when even higher accuracy is desired, 99.9% CIs are computed (the corresponding value of $\alpha$ is 0.001).

**Example.** Before we study a general theory of how CIs are constructed, we consider an example of a sample $X_1, \ldots, X_n$ drawn from normal distribution with unknown mean $\mu$ and a known variance $\sigma^2$. We would like to construct a $100(1-\alpha)\%$ CI for $\mu$. We know that $\widehat{\mu} = \bar{X}$ is the UMVUE for $\mu$. By the CLT, $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

Define by $z_{\alpha/2}$ the **critical value** (or a cut-off point) for a standard normal distribution such that $\mathbb{P}(Z > z_{\alpha/2}) = \alpha/2$ where $Z \sim N(0,1)$. For $\alpha = 0.1$ (90% CI), $z_{\alpha/2} = z_{0.05} = 1.645$; for $\alpha = 0.05$ (95% CI), $z_{\alpha/2} = z_{0.025} = 1.96$; and for $\alpha = 0.01$ (99% CI), $z_{\alpha/2} = z_{0.005} = 2.576$.

Continuing with construction of a CI, we write

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

From here,

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

This defines a $100(1-\alpha)\%$ CI for $\mu$, which covers the true $\mu$ with probability $1-\alpha$, or roughly, for $100(1-\alpha)\%$ of the samples. The CI for $\mu$ is

$$\Big[ X - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ X + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \Big],$$

or, $X \pm z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$.

Note that the margin of error of this CI is $m = z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$, and it satisfies the two properties discussed earlier. It is directly proportional to the critical value $z_{\alpha/2}$ which increases as the confidence level increases, and it is inversely proportional to the sample size $n$, so as $n$ grows, the margin of error shrinks.

**Example.** Suppose a sample of size 100 is drawn from a normally distributed population with a known standard deviation of 11.5. The sample mean is found to be 40.2. We would like to compute a 95% CI for the mean. We are given that $n = 100, \alpha = 0.05, \sigma = 11.5$, and $\bar{X} = 40.2$. We compute

$\bar{X} \pm z_{\alpha/2}\sigma/\sqrt{n} = 40.2 \pm (1.96)(11.5)/\sqrt{100} = 40.2 \pm 2.25 = [37.95,\ 42.45]$. $\square$

## CALCULATING REQUIRED SAMPLE SIZE FOR A GIVEN MARGIN OF ERROR

Assume that a population has a normal distribution with a known standard deviation $\sigma$, and suppose we would like to compute a $100(1-\alpha)\%$ CI for the mean such that its margin of error doesn't exceed a specific value $m$. How many individuals should we sample, that is, what is the required sample size $n$ that would give us the desired margin of error?

To answer this question, we proceed as follows. We are given $\sigma$, $Z_{\alpha/2}$, and the upper bound for the margin of error $m$. We need to calculate $n$. We write

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq m,$$

from where

$$n \geq \left(\frac{z_{\alpha/2}\sigma}{m}\right)^2.$$

Thus, it is sufficient to take $n$ as the smallest integer just above $\left(\frac{z_{\alpha/2}\sigma}{m}\right)^2$. It means that

$$n = \left\lceil \left(\frac{z_{\alpha/2}\sigma}{m}\right)^2 \right\rceil.$$

**Example.** In the previous example, the margin of error was 2.25. Suppose we would like the margin of error not to exceed 2. We are given $\alpha = 0.05, z_{\alpha/2} = 1.96, \sigma = 11.5$, and $m = 2$. We need to find a minimum required sample size. We compute

$$n = \left\lceil \left(\frac{z_{\alpha/2}\sigma}{m}\right)^2 \right\rceil = n = \left\lceil \left(\frac{(1.96)(11.5)}{2}\right)^2 \right\rceil = n = \lceil 127.0129 \rceil = 128.$$

For a sample of size 128, the actual margin of error will be $(1.96)(11.5)/\sqrt{128} = 1.992 < 2$. $\square$

## PIVOTAL METHOD FOR CONSTRUCTION OF CONFIDENCE INTERVALS

**Definition.** Suppose $X_1, \ldots, X_n \sim f(x|\theta)$. A **pivotal quantity** (or **pivot**) is a random variable that depends on $X_1, \ldots, X_n$ and $\theta$ which distribution doesn't depend on $\theta$.

**Example.** For $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, the random variable $\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ is a pivotal quantity. $\square$

**How to construct a CI based on a pivotal quantity?** Let $T(X_1, \ldots, X_n, \theta)$ denote a pivotal quantity. To construct a $100(1-\alpha)\%$ CI for $\theta$, we first find the cut-off points of the middle interval above which $100(1-\alpha)\%$ of the area under the density curve lies. Call these cut-offs $z_{1-\alpha/2}$ and $z_{\alpha/2}$. We have $\mathbb{P}\big(z_{1-\alpha/2} < T(X_1, \ldots, X_n, \theta) < z_{\alpha/2}\big) = 1 - \alpha$. The next step is to solve the double inequality for $\theta$ and obtain the lower and upper bounds of the confidence interval $[A, B]$: $\mathbb{P}(A < \theta < B) = 1 - \alpha$.

**Example.** In the case of normal distribution, we take the pivot $\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ and two cut-offs $z_{1-\alpha/2} = -z_{\alpha/2}$ (because of symmetry) and $z_{\alpha/2}$, and write

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

Solving for $\mu$, we get

$$\mathbb{P}\Big(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big) \;=\; 1 - \alpha,$$

which gives us the CI for $\mu$ as $\Big[\bar{X} - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}\Big]$. $\square$

**Example.** Suppose $n = 1$ and $X$ is exponentially distributed random variable with mean $\beta$. The pdf is $f_X(x) = \dfrac{1}{\beta}e^{-x/\beta},\ x > 0$, and the cdf is $F_X(x) = 1 - e^{-x/\beta},\ x > 0$. The random variable $Y = X/\beta$ is a pivotal quantity since its cdf is

$$F_Y(y) = \mathbb{P}(Y < y) = \mathbb{P}\Big(\frac{X}{\beta} < y\Big) = \mathbb{P}(X < y\beta) = F_X(y\beta) = 1 - e^{-(y\beta)/\beta} = 1 - e^{-y},$$

which doesn't depend on $\beta$. Using this pivotal quantity, we write $\mathbb{P}(a < Y < b) = 1 - \alpha$ where $a$ and $b$ are the lower and upper cut-offs for an exponential distribution with mean 1. They can be found from equations $F_Y(a) = \alpha/2$ and $F_Y(b) = 1 - \alpha/2$. Therefore, a $100(1-\alpha)\%$ CI for $\beta$ is found from the system of three equations:

$$\begin{cases} \mathbb{P}\Big(a < \frac{X}{\beta} < b\Big) = 1 - \alpha \\ 1 - e^{-a} = \frac{\alpha}{2} \\ 1 - e^{-b} = 1 - \frac{\alpha}{2}. \end{cases}$$

Solving, we obtain that $a = -\ln(1 - \alpha/2)$, $b = -\ln(\alpha/2)$, and $\mathbb{P}\Big(\dfrac{X}{b} < \beta < \dfrac{X}{a}\Big) = 1 - \alpha$. Thus, a

CI for $\beta$ is

$$\Big[\frac{X}{b},\ \frac{X}{a}\Big] = \Big[-\frac{X}{\ln(\alpha/2)},\ -\frac{X}{\ln(1-\alpha/2)}\Big].$$

For example, if the observed value of $X$ is 2.5, then a 95% CI for $\beta$ is

$$\Big[-\frac{X}{\ln(\alpha/2)},\ -\frac{X}{\ln(1-\alpha/2)}\Big] = \Big[-\frac{2.5}{\ln(0.025)},\ -\frac{2.5}{\ln(0.975)}\Big] = [0.68,\ 98.74].$$

Note how wide this interval is since it is based on a sample size of 1. □

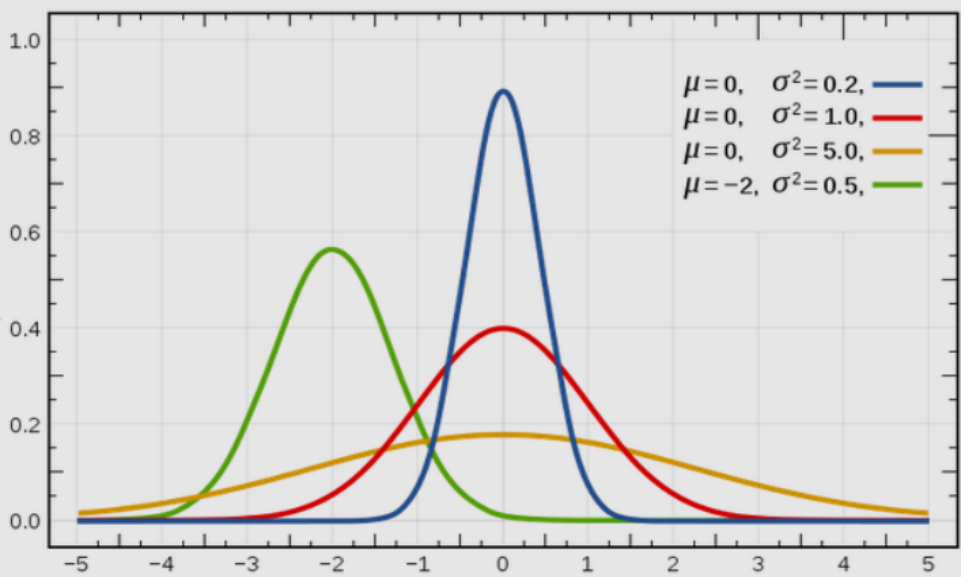
**How to determine a pivotal quantity for a given distribution?**

**Definition.** If pdf of a distribution has the form $f\Big(\frac{x-\theta_1}{\theta_2}\Big)$, then $\theta_1$ is called a **location parameter**, and $\theta_2$ is termed a **scale parameter**.

**Proposition.** If $X \sim f, F\Big(\frac{x-\theta_1}{\theta_2}\Big)$, then $Y = \dfrac{X-\theta_1}{\theta_2}$ is a pivotal quantity.

**Proof.** $F_Y(y) = \mathbb{P}(Y < y) = \mathbb{P}\Big(\dfrac{X-\theta_1}{\theta_2} < y\Big) = \mathbb{P}(X < \theta_1 + y\theta_2) = F_X(\theta_1 + y\theta_2) = F\Big(\dfrac{\theta_1 + y\theta_2 - \theta_1}{\theta_2}\Big) = F(y)$, which doesn't depend on the parameters. □

**Example.** For a $N(\mu, \sigma^2)$ distribution, the parameter $\mu$ is responsible for location of the density and $\sigma$ for the scale. The density curve slides along the $x$-axis as $\mu$ changes (thus, location), whereas as $\sigma$ changes, the density becomes more stretched out or squeezed. We illustrate it with the figure.




The variable $(X-\mu)/\sigma$ is a pivotal quantity. □

**Example.** For an exponential distribution with mean $\beta$, the density is $\frac{1}{\beta}e^{-x/\beta}$, $x>0$. Since we divide by $\beta$, it is a scale parameter and $X/\beta \sim Exp(1)$ is a pivotal quantity. □

**Example.** Consider $X \sim Unif(0,\theta)$. The density is $f_X(x)=\frac{1}{\theta}$, $0<x<\theta$. Here $\theta$ is a scale parameter, and $U=X/\theta \sim Unif(0,1)$ is a pivotal quantity. We can find a $100(1-\alpha)\%$ CI for $\theta$ by computing $\mathbb{P}\Big(a<\dfrac{X}{\theta}<b\Big)=1-\alpha$ where $\mathbb{P}(U<a)=\alpha/2$ and $\mathbb{P}(U<b)=1-\alpha/2$. From here, $a=\alpha/2$ and $b=1-\alpha/2$. Thus, the CI is $\left[\frac{X}{b},\ \frac{X}{a}\right]=\left[\frac{X}{1-\alpha/2},\ \frac{X}{\alpha/2}\right]$. If, for instance, $X=6.7$, a 95% CI for $\theta$ is $\left[\frac{6.7}{0.975},\ \frac{6.7}{0.025}\right]=[6.89, 268.00]$. Note how wide this CI is because it is based on a single observation. □

**Example.** Consider $X_1,\ldots,X_n \sim Unif(0,\theta)$. The MLE $X_{(n)} \sim \frac{nx^{n-1}}{\theta^n}$, $0<x<\theta$, where $\theta$ is a scale parameter, and thus, $X_{(n)}/\theta$ is a pivot. Its pdf is $nx^{n-1}$ and cdf is $F(x)=x^n$, for $0<x<1$, and hence a $100(1-\alpha)\%$ CI for $\theta$ is found from the following equations:

$$\mathbb{P}\Big(a<\frac{X_{(n)}}{\theta}<b\Big)=1-\alpha,\quad F(a)=a^n=\alpha/2,\quad \text{and}\quad F(b)=b^n=1-\alpha/2.$$

Consequently, the CI is

$$\Big[\frac{X_{(n)}}{b},\ \frac{X_{(n)}}{a}\Big] = \Big[\frac{X_{(n)}}{(1-\alpha/2)^{1/n}},\ \frac{X_{(n)}}{(\alpha/2)^{1/n}}\Big].$$

As a numerical example, assume that we observed $4.5, 1.8, 3.2, 2.2, 6.7$, and $5.7$. The maximum is $X_{(6)} = 6.7$, and thus, a 95% CI, say, for $\theta$ is $\Big[\frac{6.7}{(0.975)^{1/6}},\ \frac{6.7}{(0.025)^{1/6}}\Big] = [6.73,\ 12.39]$. $\square$

## CONFIDENCE INTERVAL FOR DIFFERENCE OF TWO MEANS

Suppose we have a sample of size $n_1$ drawn from a $N(\mu_1, \sigma_1^2)$ distribution, and another independent sample of size $n_2$ drawn from a $N(\mu_2, \sigma_2^2)$ distribution. We would like to compute a $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$. To this end, we notice that $\bar{X}_1 \sim N\Big(\mu_1, \frac{\sigma_1^2}{n_1}\Big)$ and $\bar{X}_2 \sim N\Big(\mu_2, \frac{\sigma_2^2}{n_2}\Big)$. Therefore,

$$\bar{X}_1 - \bar{X}_2 \sim N\Big(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\Big),$$

and thus,

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

is a pivotal quantity. The CI has the form

$$\Big[\bar{X}_1 - \bar{X}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ \bar{X}_1 - \bar{X}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\Big].$$

To show how computations work, let's assume that $\bar{X}_1 = 1.2, n_1 = 36$, and $\sigma_1 = 0.8$, and $\bar{X}_2 = 2.7, n_2 = 42$, and $\sigma_2 = 1.1$. A 99% CI for $\mu_1 - \mu_2$ is

$$1.2 - 2.7 \pm 2.576\sqrt{\frac{(0.8)^2}{36} + \frac{(1.1)^2}{42}} = -1.5 \pm 0.56 = \big[-2.06, -0.94\big]. \quad \square$$

## CALCULATING REQUIRED SAMPLE SIZE FOR A GIVEN MARGIN OF ERROR

Consider two independent samples from normal distributions and assume that variances are equal and known, that is, $\sigma_1 = \sigma_2 = \sigma$ where the value of $\sigma$ is given. Suppose we want to sample a total of $N$ individuals.

**Claim.** The variance $\mathbb{V}ar(\bar{X}_1 - \bar{X}_2)$ is minimal if $n_1 = n_2 = N/2$.

**Proof:** We assume that one sample has size $n$ and the second sample has size $N-n$. We need to minimize with respect to $n$ the variance $\mathbb{V}ar(\bar{X}_1 - \bar{X}_2) = \dfrac{\sigma^2}{n} + \dfrac{\sigma^2}{N-n}$. Taking the derivative with respect to $n$ and setting it equal to zero, we get

$$-\frac{\sigma^2}{n^2} + \frac{\sigma^2}{(N-n)^2} = 0.$$

From here, $N - n = n$ or $n = N/2$. $\square$

Further, suppose we need to find a minimal required sample size per group for a $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ with a margin of error not exceeding a pre-specified value of $m$. We have

$$z_{\alpha/2}\sqrt{\frac{\sigma^2}{N/2} + \frac{\sigma^2}{N/2}} \le m, \text{ or } \frac{2z_{\alpha/2}\sigma}{\sqrt{N}} \le m.$$

Thus,

$$N = \left\lceil \left( \frac{2z_{\alpha/2}\sigma}{m} \right)^2 \right\rceil,$$

and the minimal required sample size per sample is $\lceil N/2 \rceil$.

**Example.** To see how calculations are carried out, suppose we know that $\sigma = 1.1$ and we want the margin of error of a 95% CI for $\mu_1 - \mu_2$ to be 0.5 or smaller. We compute

$$N = \left\lceil \left( \frac{(2)(1.96)(1.1)}{0.5} \right)^2 \right\rceil = \lceil 74.4 \rceil = 75,$$

and thus, we need to sample at least $\lceil N/2 \rceil = \lceil 75/2 \rceil = 38$ individuals per sample. The actual margin of error that corresponds to this sample size is

$$1.96\sqrt{\frac{(2)(1.1)^2}{38}} = 0.495 < 0.5. \quad \square$$

## CONFIDENCE INTERVAL FOR ONE PROPORTION

Let $X$ be the number of successes in a sample of size $n$ with the probability of success $p$. We need to construct a $100(1-\alpha)\%$ CI for $p$. We know that $X \sim Bi(n,p)$. The MLE and MM estimators are $\widehat{p} = X/n$ (show it!). The mean is $\mathbb{E}\widehat{p} = \mathbb{E}(X/n) = np/n = p$ (unbiased) and the variance is $\mathbb{V}ar(\widehat{p}) = \mathbb{V}ar(X/n) = \dfrac{np(1-p)}{n^2} = \dfrac{p(1-p)}{n} \to 0$ as $n \to \infty$ (consistent). By the CLT,

$\widehat{p} \overset{appr.}{\sim} N\Big(p, \dfrac{p(1-p)}{n}\Big)$. A $100(1-\alpha)\%$ CI for $p$ is $\widehat{p} \pm z_{\alpha/2}\sqrt{\dfrac{\widehat{p}(1-\widehat{p})}{n}}$.

Suppose we draw a sample of 100 students and find that 65 of them are full-time students. We need to construct a 90% CI for the true proportion of full-time students on campus. We write $\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 0.65 \pm 1.645\sqrt{\frac{0.65(1-0.65)}{100}} = 0.65 \pm 0.078 \ = \ [0.572,\, 0.728]$. $\square$

## CALCULATING REQUIRED SAMPLE SIZE FOR A GIVEN MARGIN OF ERROR

Suppose we need to find a minimal required sample size for a $100(1-\alpha)\%$ CI for $p$ with the margin of error not exceeding $m$. We write

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

Since we don't know the value of $\widehat{p}$, we need to consider the worst case scenario, that is, we need to find the maximum value of $\widehat{p}(1-\widehat{p})$. It's an upside-down parabola that reaches its maximum in the center where $p = 1/2$. The maximum value is $1/2(1-1/2) = 1/4$. Thus, we have

$$z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \ \leq \ z_{\alpha/2}\sqrt{\frac{1/4}{n}} = \frac{z_{\alpha/2}}{2\sqrt{n}} \ \leq \ m.$$

The value $n = \left\lceil \left(\dfrac{z_{\alpha/2}}{2m}\right)^2 \right\rceil$ is called the **most conservative estimate of** $n$.

**Example.** Suppose we want to find the most conservative estimate of $n$ for a 99% CI for $p$ if we don't want the width of the interval ($2m$) to exceed 10%. We are given $z_{\alpha/2} = z_{0.005} = 2.576$, and $2m = 0.1$. We compute

$$n = \left\lceil \left(\frac{2.576}{0.1}\right)^2 \right\rceil \ = \ \lceil 663.6 \rceil = 664. \quad \square$$

## CONFIDENCE INTERVAL FOR DIFFERENCE OF TWO PROPORTIONS

Suppose $X_1 \sim Bi(n_1, p_1)$ and $X_2 \sim Bi(n_2, p_2)$. We would like to construct a $100(1-\alpha)\%$ CI for $p_1 - p_2$. We note that by the CLT,

$$\widehat{p}_1 - \widehat{p}_2 \overset{appr.}{\sim} N\Big(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\Big).$$

Hence, a $100(1-\alpha)\%$ CI is

$$\widehat{p}_1 - \widehat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}.$$

**Example.** Suppose we sample 500 freshmen and 500 sophomores, and find that 240 freshmen and 170 sophomores work full-time. We need to find a 90% CI for the difference in true population proportions. We estimate $\widehat{p}_1 = 240/500 = 0.48$ and $\widehat{p}_2 = 170/500 = 0.34$, and calculate the CI as

$$0.48 - 0.34 \pm 1.645\sqrt{\frac{(0.48)(1-0.48)}{500} + \frac{(0.34)(1-0.34)}{500}} \;=\; 0.14 \pm 0.051 \;= [0.089,\; 0.191]. \quad \square$$

## HYPOTHESES TESTING

Suppose we collect a sample and would like to verify a claim that $\mu > 5$, say. We write the **null hypothesis** $H_0: \; \mu = 5$ and the **alternative hypothesis** $H_1: \; \mu > 5$. This is a **one-sided, upper-tailed** hypothesis. There can potentially also be **lower-tailed** hypothesis $H_1: \; \mu < 5$ or **two-tailed** hypothesis $H_1: \; \mu \neq 5$.

In a word problem, the statement that we need to verify becomes our alternative hypothesis. For example, we might want to check if an average weight of 20-oz Coke bottles is less than 20 oz. ($H_1: \; \mu < 20$), or that average night stay in a hotel differs from 2.5 nights ($H_1: \; \mu \neq 2.5$).

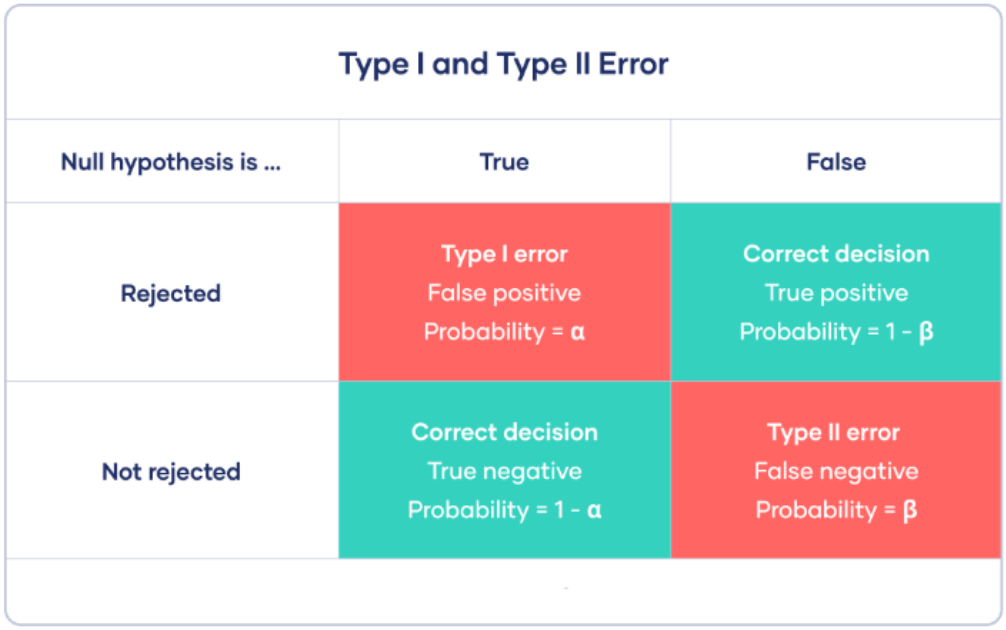
Technically speaking, the null hypothesis should complement the alternative. For example, if the claim is that $H_1: \; \mu > 5$, then the null hypothesis should be $H_0: \; \mu <= 5$. To simplify explanations(and calculations), we will always assume exact equality in null hypotheses. In our example, we will write $H_0: \; \mu = 5$.

A null hypothesis is called "null" because it essentially symbolizes null change, no improvement, no effect.

After we do some statistical testing, we make a **decision**: we either **accept** $H_1$ (at the same time **reject** $H_0$), or reject $H_1$ (at the same time accept $H_0$, or **fail to reject** $H_0$, and draw conclusion

(in plain English).

When making a decision of accepting or rejecting $H_1$, we can potentially commit either of two types of error.




**Definition. Type I error** is rejecting a true null hypothesis. It is also termed **false positive** or **false alarm**. The probability of type I error is denoted by $\alpha = \mathbb{P}(\text{reject } H_0 \,|\, H_0 \text{ is true})$.

**Definition. Type II error** is rejecting a true alternative hypothesis. It is also called **false negative** or **failure to detect**. The probability of type II error is denoted by $\beta = \mathbb{P}(\text{reject } H_1 \,|\, H_1 \text{ is true})$.

**Definition** The **power** of a test is the probability to accept a true alternative hypothesis, $1 - \beta = \mathbb{P}(\text{accept } H_1 \,|\, H_1 \text{ is true})$.

**Example.** A smoke detector in a building goes off by mistake when there is no fire. No fire means that the null hypothesis is true, but we accept the alternative hypothesis concluding that there is a fire, thus committing a type I error, commonly termed "false alarm". The other type of mistake is when there is fire but the smoke detector fails to go off. This is a type II error (failure to detect). Note that in this example, a type II error is much more serious than a type I error, since it might result in fatalities as opposed to a charge for a false call to 911.

The power of a test in this case is to detect fire when fire is actually present. To increase the power (and, respectively, decrease type II error), buildings often have several smoke detectors working independently. If, say, there are three independent smoke detectors and the probability of failing to detect a fire is 0.05 for each of them, then in combination, the probability to detect fire is $\mathbb{P}$(at least one detector works) $= 1 - \mathbb{P}$(all three detectors fail) $= 1 - \mathbb{P}(\text{detector fails})^3 = 1 - 0.05^3 = 0.999875$ as opposed to 0.95, if there only one smoke detector. $\square$

**Example.** Suppose a new medical device can detect a certain type of cancer. A type I error (false alarm) is when the device diagnoses cancer in a cancer-free patient, and a type II error (failure to detect) is when the device fails to diagnose cancer in a patient with cancer. Note that here again type I error is much milder than type II error which will result in the death of an untreated cancer patient. Type I error will probably result in unnecessary chemotherapy of a healthy patient, which in itself is not very good but most likely not fatal. Note that all medical devices are required to have a very small probability of type II error, specifically, because it results in death of a patient. $\square$

**Note.** Why can't we have a smoke detector or a medical device that has both $\alpha$ and $\beta$, probabilities of type I and type II errors, equal to zero? As we will show later in this course, $\alpha$ and $\beta$ work in opposite directions. As $\alpha$ decreases, $\beta$ increases. So, it is not possible to make them both equal to zero. The relation between $\alpha$ and $\beta$ is quite complicated even in the simplest case of the normal underlying distribution. Derivation of this relation is forthcoming.

**Definition.** Another name for the probability of type I error, $\alpha$, is the **significance level of a test** (or **level of significance**). In hypotheses testing, $\alpha$ is typically taken as 0.05 or 0.01. If not specified, we will assume $\alpha = 0.05$.

**Note.** Recall that when we compute confidence intervals, we also use $\alpha$. This is not a coincidence. Later we will show that it is the same $\alpha$. $\square$

### Explanation of how hypotheses testing works and why

Suppose we observe a sample of size $n = 49$ for which $\bar{x} = 12.2$. Assume that $\sigma$ is known to be 0.7. We would like to test that the true population mean exceeds 12 at the 5% significance level (that is, $\alpha = 0.05$). Note that the sample mean exceeds 12, but it might be due to chance, and the true $\mu$ is, in fact, less than 12.

We want to test $H_0: \ \mu = 12$ against $H_1: \ \mu > 12$. By the CLT, $\bar{X} \overset{appr.}{\sim} N(\mu, \sigma^2/n)$. We will assume that the null is true. Under $H_0$, $\bar{x} \sim N(12, 0.7^2/49)$. How likely is it to observe $\bar{x} = 12.2$, if $\bar{x}$ has this distribution? We can tell how unusual this observation is under the null hypothesis by computing the probability to fall above the observed value, $\mathbb{P}(\bar{x} > 12.2) = \mathbb{P}\big(Z > \dfrac{12.2 - 12}{0.7/\sqrt{49}}\big) = \mathbb{P}(Z > 2) = 0.0228$.

This is not a large probability, so $H_0$ is likely not to be true, and we should be leaning more towards accepting $H_1$. The decision should be to accept $H_1$ and conclude that the claim is true.

## STEPS TO CONDUCT HYPOTHESES TESTING

Step 1. Write down $H_0: \ \theta = \theta_0$ vs. $H_1: \ \theta \geq \theta_0$ (upper-tailed), or $H_1: \ \theta \leq \theta_0$ (lower-tailed), or $H_1: \ \theta \neq \theta_0$ (two-sided).

Step 2. Write down all given quantities, including $\alpha$ (use 0.05 by default).

Step 3. Compute **test statistic** and specify its distribution under the null hypothesis. In the above example, the test statistic is $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$, which under $H_0: \ \mu = \mu_0$ has a $N(0,1)$ distribution.

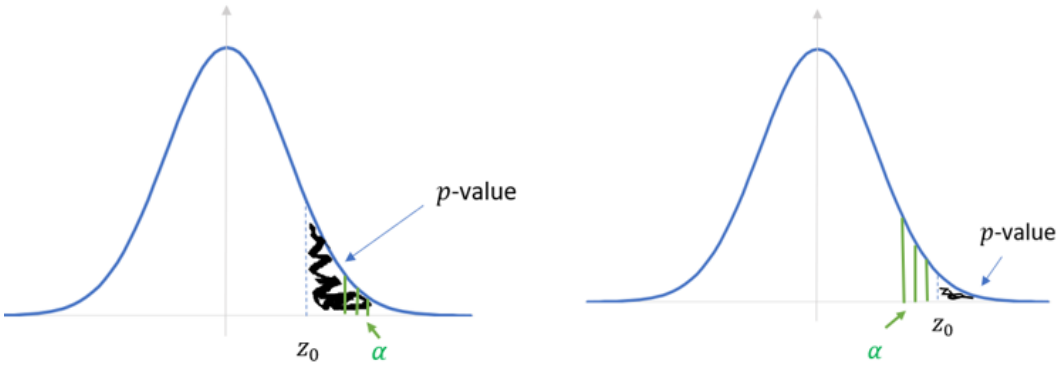
Step 4. Compute the ***p*-value** that measures how unusual the observed test statistic is assuming the null is true. The formal rule to compute $p$-value is as follows:

If $H_1: \ \theta > \theta_0$, then $p$-value$= \mathbb{P}(Z > z_0)$ where $z_0$ denotes the observed test statistic, and $Z$ is the random variable that has the same distribution as the test statistic under $H_0$.

If $H_1: \ \theta < \theta_0$, then $p$-value$= \mathbb{P}(Z < z_0)$. Note that $Z_0$ is necessarily negative.

If $H_1: \ \theta \neq \theta_0$, then $p$-value$= \mathbb{P}(Z > |z_0| \text{ or } Z < -|z_0|) = \mathbb{P}(Z > |z_0|) + \mathbb{P}(Z < -|z_0|) = 2\mathbb{P}(Z > |z_0|)$, if the underlying distribution is symmetric.

Step 5. Compare the $p$-value with significance level $\alpha$.




If $p$-value$> \alpha$, then the observed test statistic is a usual observation under $H_0$, so the null should not be rejected. If, however, $p$-value$< \alpha$, then the observed test statistic lies way out in the tail

and should be considered an unusual observation under $H_0$, so we should reject $H_0$ in favor of $H_1$.

Step 6. Once we state our decision, we need to write the conclusion in plain English (using the claim as the guideline for terminology). Conclusions are written in non-technical language for clients. As an example, we might want to say "There is enough evidence in the data to conclude that the new medical device is efficient".

## ONE-SAMPLE z-TEST FOR $\mu$

Step 1. Identify $\bar{x}$, $\sigma$, $n \leq 30$, $\mu_0$ and $\alpha$.
Step 2. Write down the hypotheses: $H_0: \ \mu = \mu_0$ vs. $H_1: \ \mu \leq \mu_0$, or $\mu \leq \mu_0$, or $\mu \neq \mu_0$.
Step 3. Compute the test statistic $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$.
Step 4. Compute the $p$-value: (i) if $H_1: \ \mu \geq \mu_0$, $p$-value$= \mathbb{P}(Z > z)$ where $Z \sim N(0,1)$; (ii) if $H_1: \ \mu \leq \mu_0$, $p$-value$= \mathbb{P}(Z < z)$; and (iii) if $H_1: \ \mu \neq \mu_0$, $p$-value$= 2\mathbb{P}(Z > |z|)$.
Step 5. Compare the $p$-value to $\alpha$ and make a decision: (i) if $p$-value$> \alpha$, fail to reject $H_0$; (ii) if $p$-value$< \alpha$, reject $H_0$.
Step 6. State the conclusion in a simple non-technical language.

**Example.** In our previous example, $\bar{x} = 12.2, \sigma = 0.7, n = 49, \mu_0 = 12$, and $\alpha = 0.05$. We test $H_0: \ \mu = 12$ against $H_1: \ \mu \geq 12$ (upper-tailed test). We compute the test statistic $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{12.2 - 12}{0.7/\sqrt{49}} = 2$. The test statistic has $N(0,1)$ distribution under $H_0$. The $p$-value$= \mathbb{P}(Z > 2) = 0.0228 < 0.05 = \alpha$. So, we reject $H_0$ and conclude that there is sufficient evidence to support the claim that $\mu > 12$. □

**Example.** Suppose we have a machine that fills bottles with 20oz of soda. We randomly collect 100 bottles and find that $\bar{x} = 19.85$. We assume that $\sigma$ is known to be 1oz. We also have $n = 100$. We want to test $H_0: \ \mu = 20$ vs. $H_1: \ \mu < 20$ (lower-tailed test). We compute the test statistic $z = \dfrac{19.85 - 20}{1/\sqrt{100}} = -1.5$. The $p$-value$= \mathbb{P}(Z < -1.5) = \mathbb{P}(Z > 1.5) = 0.0668 > 0.05 = \alpha$. So, we fail to reject $H_0$ and conclude that there is not enough evidence in the data to support the claim that the machine fills bottles with less than 20oz of soda (that is, the machine is working properly). □

**Example.** Bus transportation claims that the mean wait time for a bus is 5 minutes. We would like to test at the 1% significance level the claim that the mean wait time differs from 5 minutes. Suppose we record wait times for 36 buses and observe a mean wait of 5.9 minutes. We also assume that the standard deviation is known to be 2 minutes. We test $H_0: \ \mu = 5$ against $H_1: \ \mu \neq 5$. We

are given $\bar{x} = 6.9, \sigma = 1.2, n = 36, \mu_0 = 5$, and $\alpha = 0.01$. The test statistic is $z = \dfrac{5.9 - 5}{2/\sqrt{36}} = 2.7$. The $p$-value$= 2\mathbb{P}(Z > 2.7) = 0.0069 < 0.01$, thus, we reject the null and conclude that the average wait time differs from 5 minutes. $\square$

## REJECTION REGION AS AN ALTERNATIVE TO $P$-VALUE

Instead of computing $p$-value$= \mathbb{P}(Z > z)$, and comparing it to an $\alpha$, we can compute the critical value $z_\alpha$ corresponding to $\alpha$, that is, $\mathbb{P}(Z > z_\alpha) = \alpha$, and compare the observed test statistic $z_0$ to the critical value $z_\alpha$. If $z_0 < z_\alpha$, we see that the observation is not unusual under $H_0$ and so we don't reject the null hypothesis. If, however, $z_0 \geq z_\alpha$, then we observed something way in the tail and so we reject $H_0$.

**Definition. Rejection region** is a set of all values of the test statistic for which the null hypothesis is rejected. The complement of the rejection region is termed the **acceptance region**. The rejection region is denoted by $RR$, and the acceptance region is denoted by $AR$.

- If $H_1: \ \mu \, \mu_0$ (upper-tailed), $RR = \{z: \ z > z_\alpha\}$.
- If $H_1: \ \mu < \mu_0$ (lower-tailed), $RR = \{z: \ z < -z_\alpha\}$.
- If $H_1: \ \mu \neq \mu_0$ (two-tailed), $RR = \{z: \ z < -z_{\alpha/2} \text{ or } z > z_{\alpha/2}\} = \{z: \ |z| > z_{\alpha/2}\}$.

## CONNECTION BETWEEN CONFIDENCE INTERVAL AND HYPOTHESES TESTING

Hypotheses testing can be conducted by computing the appropriate confidence interval. Here is the explanation:

- If $H_1: \ \mu > \mu_0$, then $RR = \{z: \ z > z_\alpha\} = \left\{\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \ > \ z_\alpha\right\} = \left\{\mu_0 \ < \ \bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}\right\}$. It means that we can construct a $100(1-\alpha)\%$ one-sided confidence interval $\left[\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}, \ \infty\right)$, and, if $\mu_0$ is below this interval, we reject $H_0$ (and accept $H_1$).

- If $H_1: \ \mu < \mu_0$, then $RR = \{z: \ z < -z_\alpha\} = \left\{\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \ < \ -z_\alpha\right\} = \left\{\mu_0 \ > \ \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}\right\}$. It means that we can construct a $100(1-\alpha)\%$ one-sided confidence interval $\left(-\infty, \ \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}\right]$, and, if $\mu_0$ is above this interval, we reject $H_0$ (and accept $H_1$).

- If $H_1: \ \mu \neq \mu_0$, then $RR = \{z: \ |z| > z_{\alpha/2}\} = \left\{\left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right| \ > \ z_{\alpha/2}\right\} = \left\{\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \ < \ -z_{\alpha/2} \text{ or}\right.$

$\frac{\bar{x}-\mu_0}{\sigma\sqrt{n}} > z_{\alpha/2}\Big\} = \Big\{\mu_0 > \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ or $\mu_0 < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big\}$. It means that we can construct a regular, two-sided $100(1-\alpha)\%$ confidence interval $\Big[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big]$, and, if $\mu_0$ is not covered by this interval, we reject $H_0$ (and accept $H_1$).

## RELATION BETWEEN $\alpha$ AND $\beta$ FOR $z$-TEST FOR $\mu$, WHEN $\sigma$ IS KNOWN

Below we derive the formula that relates the probability of type I error $\alpha$ and the probability of type II error $\beta$ for the simplest case of a one-sample $z$-test for the mean $\mu$ when standard deviation $\sigma$ is known. We write $\alpha = \mathbb{P}(\text{reject } H_0 \mid H_0 \text{ is true}) = \mathbb{P}(z \in RR \,|\, H_0 \text{ is true}) = \mathbb{P}(Z > k \mid Z \sim N(0,1))$. Therefore, the critical value for the rejection region $k = \Phi^{-1}(1-\alpha) = z_\alpha$.

Turning now to $\beta$, we write $\beta = \mathbb{P}(\text{not reject } H_0 \mid H_1 \text{ is true}) = \mathbb{P}(z \in AR \mid H_1 \text{ is true}) = \mathbb{P}(z < z_\alpha \mid H_1 :\ \mu > \mu_0)$. Next, even if we know that $\mu > \mu_0$ holds, we still need to provide a specific value of $\mu$. We denote by $\delta = \mu - \mu_0$ the **effect size**. Note that under $H_1$, the specific value of $\mu$ is $\delta + \mu_0$. We can then continue,

$$\begin{aligned}
\beta &= \mathbb{P}\Big(Z < z_\alpha \mid Z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}},\ \bar{x} \sim N(\delta+\mu_0, \sigma^2/n)\Big) \\
&= \mathbb{P}\Big(\bar{x} < \mu_0 + z_\alpha\frac{\sigma}{\sqrt{n}} \mid \bar{x} \sim N(\delta+\mu_0, \sigma^2/n)\Big) \\
&= \mathbb{P}\Big(Z < \frac{\mu_0 + z_\alpha\frac{\sigma}{\sqrt{n}} - (\delta+\mu_0)}{\sigma/\sqrt{n}}\Big) = \mathbb{P}\Big(Z < z_\alpha - \frac{\delta}{\sigma}\sqrt{n}\Big) \\
&= \Phi\Big(z_\alpha - \frac{\delta}{\sigma}\sqrt{n}\Big) = \Phi\Big(\Phi^{-1}(1-\alpha) - \frac{\delta}{\sigma}\sqrt{n}\Big).
\end{aligned}$$

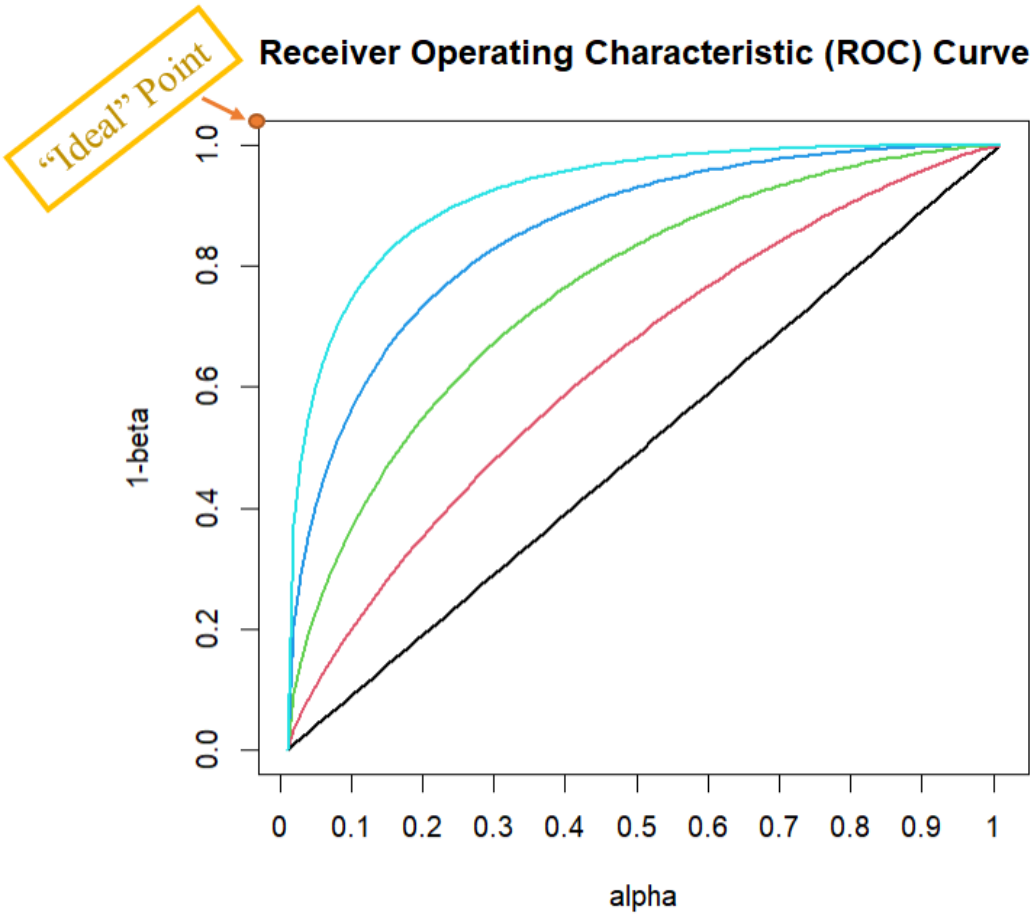
Note that if the effect size $\delta$ is close to zero, $\beta \approx \Phi(\Phi^{-1}(1-\alpha)) = 1-\alpha$, so if we want to decrease $\alpha$, $\beta$ will increase. On the other hand, if $\delta$ is very large, then $\beta$ goes to zero. It is an intuitive result, because a large effect size means that, for instance, a medical device is very efficient, so it is almost without an error we will accept the alternative and market the device.

## RECEIVER OPERATING CHARACTERISTIC CURVE

**Definition.** The **Receiver Operating Characteristic (ROC) Curve** is the plot of power$= 1-\beta = 1 - \Phi\Big(\Phi^{-1}(1-\alpha) - \frac{\delta}{\sigma}\sqrt{n}\Big)$ against $\alpha$ for different values of $\frac{\delta}{\sigma}\sqrt{n}$.

In this coordinate system, there is an "ideal" point (0,1), where $\alpha = 0$ and power$= 1$ (or, equivalently, $\beta = 0$). This point is unattainable in practice. As depicted in the figure, for $\frac{\delta}{\sigma}\sqrt{n} = 0$, it

is just the bisector *power* $= \alpha$, and once this quantity increases, the ROC curve become more and more convex, getting closer and closer to the "ideal" point.




ROC curves are used by quality control engineers to find appropriate sample size $n$ that corresponds to fixed $\alpha$, $\beta$, and $\delta/\sigma$, or to find $\delta/\sigma$ for fixed $\alpha, \beta$, and $n$.

## ONE-SAMPLE TEST FOR POPULATION PROPORTION

Suppose a sample of size $n (n \leq 30)$ is drawn and the sample proportion $\widehat{p}$ is observed. We want to test $H_0 : \ p = p_0$ against $H_1 : \ p > p_0$ or $p < p_0$ or $p \neq p_0$ where $p$ is the true population proportion. By the CLT, an approximate distribution of $\widehat{p}$ is normal with mean $p$ and standard deviation $\sqrt{\dfrac{p(1-p)}{n}}$. Thus, under $H_0 : \ p = p_0$, the approximate distribution is normal with mean

$p_0$ and standard deviation $\sqrt{\dfrac{p_0(1-p_0)}{n}}$. And the test statistic is

$$z = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

which under $H_0$ has approximately a standard normal distribution.

**Example.** In a random sample of 100 voters, 63% said that they will vote in favor of a proposal.Can it be claimed at the 1% significance level that a majority of voters favor the proposal?

We test $H_0: \ p = 0.5$ against $H_1: \ p > 0.5$. We are given $n = 100, \widehat{p} = 0.63$, and $\alpha = 0.01$. The test statistic is

$$z = \frac{0.63 - 0.5}{\sqrt{\frac{(0.5(1-0.5)}{100}}} = 2.6.$$

The $p$-value$= \mathbb{P}(Z > 2.6) = 0.0047 < 0.01 = \alpha$, thus we reject $H_0$ and conclude that the data support the claim that a majority of voters favor the proposal. □

**Example.** To check whether a machine is producing fewer than 4% defective items, a random sample of size 200 is drawn, and 6 items are found to be defective. We need to test $H_0: \ p = 0.04$ vs. $H_1: \ p < 0.04$. We are given $n = 200, \widehat{p} = 6/200 = 0.03$, $p_0 = 0.04$, and $\alpha = 0.05$ (by default). The test statistic is $z = \dfrac{0.03 - 0.04}{\sqrt{\frac{0.04(1-0.04)}{200}}} = -0.72$. The $p$-value$= \mathbb{P}(Z < -0.72) = 0.2358 > 0.05 = \alpha$. Hence, we fail to reject $H_0$ and conclude that there is not enough evidence to state that the machine produces fewer than 4% defective items. □

**Example.** A transportation authority claims that 80% of all bus trips are as-scheduled. To verify this claim, a random sample of 60 bus trips was drawn and 68% of the trips were found to be as-scheduled. We want to test $H_0: \ p = 0.8$ against $H_1: \ p \neq 0.8$. We know that $n = 60, \widehat{p} = 0.68, p_0 = 0.8$ and $\alpha = 0.05$ since it is not given. We compute the test statistic $z = \dfrac{0.68 - 0.8}{\sqrt{\frac{0.8(1-0.8)}{60}}} = -2.32$. The $p$-value$= 2\mathbb{P}(Z < -2.32) = (2)(0.0102) = 0.0204 < 0.05 = \alpha$. Thus, we reject $H_0$ and conclude that the proportion of on-schedule bus trips differs from 80%. □

## TWO-SAMPLE TEST FOR TWO POPULATION PROPORTIONS

Suppose two random samples of sizes $n_1$ and $n_2$ are drawn from two independent populations. We assume that the sample sizes are comparable in magnitude and are both larger than 30. Let $x_1$ and $x_2$ denote the number of objects of interest in these samples, and let $\widehat{p}_1 = \dfrac{x_1}{n_1}$ and $\widehat{p}_2 = \dfrac{x_2}{n_2}$ be the two sample proportions. Denoteby $p_1$ and $p_2$ the true unknown population proportions. Suppose we want to test $H_0 : \ p_1 = p_2$ againt $H_1 : \ p_1 > p_2$, or $p_1 < p_2$, or $p_1 \neq p_2$. Under $H_0$, the two population proportions are equal. To estimate this common proportion $p_1 = p_2 = p$, we pool the samples to obtain the **pooled estimate**

$$\widehat{p} = \frac{x_1 + x_2}{n_1 + n_2}.$$

By the CLT,$\widehat{p}_1$ is approximately normally distributed with mean $p$ and variance $\dfrac{p(1-p)}{n_1}$. Similarly, $\widehat{p}_2$ is approximately normal with mean $p$ and variance $\dfrac{p(1-p)}{n_2}$. Thus, the difference $\widehat{p}_1 - \widehat{p}_2$ is approximately normal with mean $p - p = 0$ and variance $\dfrac{p(1-p)}{n_1} + \dfrac{p(1-p)}{n_2} = p(1-p)\Big(\dfrac{1}{n_1} + \dfrac{1}{n_2}\Big)$. Therefore, the test statistic is

$$z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)}}$$

which under $H_0$ has approximately a standard normal distribution.

**Example.** To test whether vitamin C is a preventive measure for the common cold, 500 people took vitamin C, and 500 people took a sugar pill (placebo). In the first group, 200 people had a cold, while in the second group, 230 had a cold. Suppose we would like to test the claim at the 1% significance level. We are given: $n_1 = n_2 = 500, x_1 = 200, x_2 = 230, \widehat{p}_1 = x_1/n_1 = 200/500 = 0.40, \widehat{p}_2 = x_2/n_2 = 230/500 = 0.46$, and $\alpha = 0.01$. We need to test $H_0 : \ p_1 = p_2$ against $H_1 : \ p_1 < p_2$. We calculate the pooled estimate $\widehat{p} = (200{+}230)/(500{+}500) = 430/1000 = 0.43$, and so the test statistic is

$$z = \frac{0.40 - 0.46}{\sqrt{0.43(1-0.43)\Big(\frac{1}{500} + \frac{1}{500}\Big)}} = -1.92.$$

The $p$-value is $\mathbb{P}(Z < -1.92) = 0.0274 > 0.01$, hence we fail to reject $H_0$ at the 1% level of significance and conclude that vitamin C is not a preventive measure for the common cold. $\square$

**Example.** In a sample of 70 seniors, 30% have a job, whereas in a sample of 50 freshmen, 12% have a job. Suppose we would like to test a claim that a higher proportion of seniors have a job than

freshmen. We need to test $H_0: p_1 = p_2$ against $H_1: p_1 > p_2$. We are given $\widehat{p}_1 = 0.30, \widehat{p}_2 = 0.12$, and $\alpha = 0.05$. The pooled estimate is

$$\widehat{p} = \frac{(0.30)(70) + (0.12)(50)}{70 + 50} = \frac{21 + 6}{120} = \frac{27}{120} = 0.225.$$

The test statistic is computed as

$$z = \frac{0.30 - 0.12}{\sqrt{(0.225)(1 - 0.225)\left(\frac{1}{70} + \frac{1}{50}\right)}} = 2.328.$$

The $p$-value is $\mathbb{P}(Z > 2.328) = 0.009956 < 0.05$. Thus, we reject $H_0$ and conclude that a higher proportion of seniors have a job than freshmen. $\square$

**Example.** A survey of moviegoers during morning hours and evening hours revealed that 78 out of 100 morning visitors buy popcorn, and 94 out of 100 evening visitors buy popcorn. The manager would like to know if the true proportions differ. We test $H_0: p_1 = p_2$ vs. $H_1: p_1 \neq p_2$. Given $\widehat{p}_1 = 78/100 = 0.78, \widehat{p}_2 = 94/100 = 0.94$, and $\alpha = 0.05$. The pooled estimate of the common proportion under the null is

$$\widehat{p} = \frac{78 + 94}{100 + 100} = \frac{172}{200} = 0.86.$$

The test statistic is

$$z = \frac{0.78 - 0.94}{\sqrt{0.86(1 - 0.86)\left(\frac{1}{100} + \frac{1}{100}\right)}} = -3.2606.$$

The $p$-value$= 2\mathbb{P}(Z < -3.2606) = (2)(0.0006) = 0.0012 < 0.05$, so we reject $H_0$ and conclude that the proportions differ. $\square$

## THE T-DISTRIBUTION

**Definition.** The $t$-distribution has density function

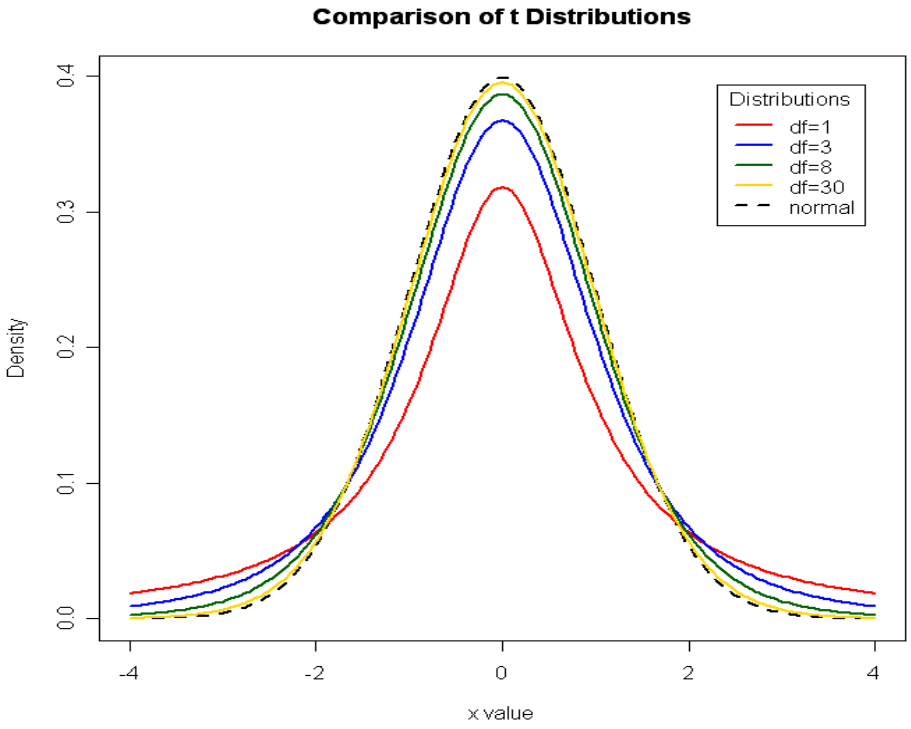
$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma\left(\frac{k}{2}\right)}\left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}, \quad -\infty < x < \infty,$$

where the parameter $k$ is called the **number of degrees of freedom** (or, simply, **degrees of freedom**).

**Note.** The number of degrees of freedom (df) is the number of choices minus the number of constraints that the choices must satisfy. For example, if we can choose any 5 numbers, then the number of degrees of freedom is 5. But if we want these numbers to add up to 100, we are free to choose 4 numbers and calculate the last one by subtraction from 100, so the number of degrees of freedom is 4 (5 numbers minus one constraint). □

**Note.** William Sealy Gosset (1876-1937) was an English statistician. He published under the pen name Student, and developed the Student's $t$-distribution in 1908.

**Note.** The density of $t$-distribution is defined on the entire real line. It is bell-shaped and symmetric around zero. As the number of degrees of freedom increases, the distribution approaches the standard normal distribution. For smallers values of $df$, the distribution has heavier tails (that is, more probabilities are in the tails, as illustrated below).




## HYPOTHESES TESTS ABOUT MEAN WHEN VARIANCE IS UNKNOWN

Suppose we are given a sample of size $n$, where either $n \geq 30$ and the CLT can be applied, or $n < 30$ but the distribution is known to be normal. Assume also that the sample mean $\bar{x}$ and

sample standard deviation $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$ are available. We need to test $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$ or $\mu < \mu_0$ or $\mu \neq \mu_0$. The test statistic is $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ which under $H_0$ has a $t$-distribution with $n-1$ degrees of freedom.

**Note.** The number of degrees of freedom is $n-1$ because there are $n$ values $x_1 - \bar{x}$ through $x_n - \bar{x}$, but there is one constraint $\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = 0$.

**Example.** A company that manufactures light bulbs claims that their light bulbs last, on average, at least 1100 hours. A sample of 25 light bulbs gave a mean life of 1160 hours with a sample standard deviation of 85 hours. Suppose we need to test the claim at the 1% significance level. Since $n < 30$, we need to make an additional assumption that the life of a bulb is normally distributed. We are testing $H_0 : \mu = 1100$ against $H_1 : \mu >= 1100$. We are given that $n = 25, \bar{x} = 1160, s = 85, \mu_0 = 1100$, and $\alpha = 0.05$. We compute the test statistic

$$t = \frac{1160 - 1100}{85/\sqrt{25}} = 3.529.$$

The number of degrees of freedom is $df = 25 - 1 = 24$. The $p$-value$= \mathbb{P}(T > 3.529) = 0.000857$ (by Excel, after typing "=1-t.dist(3.529,24,true)"). Alternatively, using the table for $t$-distribution, we can find the bounds for the $p$-value:

$$0.001 > \mathbb{P}(T > 3.529) > 0.0005.$$

We see that $p$-value$< 0.01$, thus, we reject the null hypothesis and conclude that the population mean life of a light bulb is at least 1100 hours. □

**Example.** A researcher wants to test if the mean annual salary of all lawyers in a city is less than \$125,000. A sample of 45 lawyers reveals a sample mean annual salary \$108,400 and a sample standard deviation of \$34,700. Test the hypothesis at the 5% level. We would like to test $H_0 : \mu = 125,000$ against $H_1 : \mu < 125,000$. We have that $n = 45, \bar{x} = 108,400, s = 34,700, \mu_0 = 125,000$, and $\alpha = 0.05$. The test statistic is

$$t = \frac{108,400 - 125,000}{34,700/\sqrt{45}} = -3.209,$$

which by the CLT, under the null hypothesis, has approximately $t$-distribution with $45 - 1 = 44$ degrees of freedom. From Excel, the $p$-value is $\mathbb{P}(T < -3.209) = \mathbb{P}(T > 3.209) = 0.001243 < 0.05$. If we use the table for $t$-distribution, we get that $p$-value$< 0.005 < 0.05$. Hence, we reject $H_0$ and conclude that there is enough evidence in the data to support the researchers' hypothesis. □

**Example.** The mean flight delay was 15 minutes before the merger of the two companies. The CEO wants to find out whether the mean flight delay has changed since the merger. A random sample of size 81 flights is drawn, and it is obtained that the sample average delay time is 20 minutes with a sample standard deviation of 17 minutes. To test the claim, we state $H_0: \ \mu = 15$ against $H_1: \ \mu \neq 15$. We have $n = 81, \bar{x} = 24, s = 17, \mu_0 = 15$ and $\alpha = 0.05$. The test statistic is

$$t = \frac{24 - 15}{17/\sqrt{81}} = 2.6471.$$

The number of degrees of freedom is $81 - 1 = 80$, and the $p$-value$= 2\mathbb{P}(T > 2.6471) = 0.009776 < 0.05$. Thus, we reject the null hypothesis and conclude that the mean flight delay has changed since the merger. □

## CONFIDENCE INTERVAL FOR MEAN WHEN VARIANCE IS UNKNOWN

**Definition.** A $100(1-\alpha)\%$ confidence interval for $\mu$ when $\sigma$ is not known is

$$\bar{x} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}$$

where the critical value $t_{\alpha/2,n-1}$ satisfies $\mathbb{P}(T > t_{\alpha/2,n-1}) = \alpha/2$ with $T \sim t(df = n-1)$.

**Example.** In the previous example, we conducted a two-sided hypotheses testing, which, as we know can be equivalently carried out by constructing a two-sided confidence interval. We have the quantities: $\bar{x} = 20, s = 17, n = 81$, and $\alpha = 0.05$. We compute the critical value in Excel by typing "=t.inv(0.975,80)". We get $t_{0.025,80} = 1.99$. Now we are ready to construct a 95% CI for $\mu$. We write

$$\bar{x} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}} = 20 \pm 1.99 \frac{17}{\sqrt{81}} = [16.2, 23.8].$$

We see that $\mu_0 = 15$ is outside of this interval, so the decision is that the null hypothesis should be rejected in favor of the alternative, which is in agreement with the earlier decision. □

## HYPOTHESES TEST FOR TWO MEANS BASED ON T-DISTRIBUTION

Suppose we have two independent samples drawn from two populations. The sample sizes and statistics are $n_1, \bar{x}_1, s_1, n_2, \bar{x}_2$, and $s_2$. We further assume that the sample sizes $n_1$ and $n_2$ are of comparable magnitudes, and either both are 30 or above (and so the CLT applies) or both are less than 30 but the populations are known to be normally distributed. Suppose we want to test $H_0: \ \mu_1 = \mu_2$ against $H_1: \mu_1 \geq \mu_2$ or $\mu_1 \leq \mu_2$ or $\mu_1 \neq \mu_2$. Two cases are distinguished:

• It is assumed that the population standard deviations (equivalently, variances) are equal, that is, $\sigma_1 = \sigma_2$.

• It is assumed that the population standard deviations (or variances) are not equal, that is, $\sigma_1 \neq \sigma_2$.

**Note.** The assumption of equal variances is valid typically only if samples are drawn from the same population before and after some intervention. It is believed that the intervention can only change the mean but not the variance. In all other cases, when samples are drawn from two different and independent populations, the equality of variances cannot be assumed.

### Hypotheses Test for Two Means when Standard Deviations are Equal

First, we estimate the common standard deviation by the **pooled estimate**

$$s_p = \sqrt{\frac{\sum(x-\bar{x}_1)^2 + \sum(x-\bar{x}_2)^2}{n_1 - 1 + n_2 - 1}} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}.$$

**Note.** The pooled estimate $s_p$ necessarily falls between $s_1$ and $s_2$. This follows from the observation that $s_p^2$ is the weighted average of $s_1^2$ and $s_2^2$. This fact can be used to check that the calculated value of $s_p$ lies within the bounds and thus is reasonable.

Next, we compute the test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

which under $H_0$, has a $t$-distribution with $df = n_1 + n_2 - 2$.

**Example.** A hotel manager wants to check whether advertisement increases average hotel stay. He picks at random a sample of 10 visitors before an advertisement takes place and finds that the mean stay is 2.2 nights with a standard deviation of 0.9 nights. Post advertisement he selects another random sample of size 10 and finds that the mean stay is 2.9 nights and standard deviation is 1.1 nights. We test $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 < \mu_2$. We will assume that the advertisement can influence only the mean stay and not the variance, so we will assume that $\sigma_1 = \sigma_2$. We are given that $n_1 = n_2 = 10, \bar{x}_1 = 2.2, s_1 = 0.9, \bar{x}_2 = 2.9, s_2 = 1.1$, and $\alpha = 0.05$. The pooled estimate of common standard deviation is

$$s_p = \sqrt{\frac{(10)(0.9)^2 + (10)(1.1)^2}{10+10-2}} = 1.005.$$

The test statistic is

$$t = \frac{2.2 - 2.9}{1.005\sqrt{\frac{1}{10} + \frac{1}{10}}} = -1.557.$$

The number of degrees of freedom is $df = 10 + 10 - 2 = 18$. The $p$-value$= \mathbb{P}(T < -1.557) = \mathbb{P}(T > 1.557) > 0.05 = \alpha$ (from Excel, $p$-value$= 0.0684$). We fail to reject the null hypothesis and conclude that the advertisement didn't increase average hotel stay. □

**Example.** Cholesterol levels are measured for 28 heart attack patients (case-patients) and 30 other hospital patients who didn't have a heart attack (control-patients). The sample quantities are:

| | Mean | Stdev |
|---|---|---|
| Case Group | 233.7 | 56.3 |
| Control Group | 184.2 | 49.8 |

We want to test at $\alpha = 0.001$ that the mean cholesterol level is higher for the case patients. For the analysis, we will make an assumption that the population standard deviations are equal. We test $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 > \mu_2$. We compute the pooled estimate

$$s_p = \sqrt{\frac{(28-1)(56.3)^2 + (30-1)(49.8)^2}{28 + 30 - 2}} = 53.03.$$

The test statistic is

$$t = \frac{233.7 - 184.2}{53.03\sqrt{\frac{1}{28} + \frac{1}{30}}} = 3.55.$$

The number of degrees of freedom is $df = 28 + 30 - 2 = 56$. The $p$-value$= \mathbb{P}(T > 3.55) < 0.0005 < 0.001 = \alpha$ (from Excel, $p$-value$= 0.000394$. Therefore, we reject $H_0$ and conclude that the mean cholesterol level is higher for heart attack patients. □

**Example.** A company is interested in finding out whether mean customer satisfaction scores are the same for two stores owned by this company. The data for two random samples of sizes 5 and 7 are available, with the respective sample means of 30 and 24, and sample standard deviations of 3.4 and 5.1. We carry out the two-sided test with $H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 \neq \mu_2$. We assume equal standard deviations and a significance level of 0.05. We also need to assume the normality of measurements since the sample sizes are small. We compute the pooled estimate of the standard deviation as follows.

$$s_p = \sqrt{\frac{(5-1)(3.4)^2 + (7-1)(5.1)^2}{5 + 7 - 2}} = 4.498.$$

The test statistic is

$$t = \frac{30 - 24}{4.498\sqrt{\frac{1}{5} + \frac{1}{7}}} = 2.278.$$

The number of degrees of freedom is $df = 5 + 7 - 2 = 10$, and the $p$-value$= 2\mathbb{P}(T > 2.278) < (2)(0.025) = 0.05 = \alpha$ (from Excel, $p$-value$= 0.046$). Hence, we reject $H_0$ and conclude that the mean customer satisfaction scores differ for the two stores. $\square$

### Confidence Interval for Difference in Means when Standard Deviations are Equal

**Definition.** A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ when $\sigma_1 = \sigma_2$ is

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2,df}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where the number of degrees of freedom is $df = n_1 + n_2 - 2$.

**Example.** In the previous example, we computed that $s_p = 4.498$ and $df = 10$. The critical value is $t_{0.025,10} = 2.228$. Hence, a 95% CI for $\mu_1 - \mu_2$ is

$$30 - 24 \pm (2.228)(4.498)\sqrt{\frac{1}{5} + \frac{1}{7}} \;=\; 6 \pm 5.9 \;=\; [0.1, 11.9].$$

Since the interval doesn't cover 0, we would decide in favor of the alternative, as we did in the hypotheses testing. $\square$

### Hypotheses Test for Two Means when Standard Deviations are Unequal

Suppose we want to test $H_0:\ \mu_1 = \mu_2$ against $H_1 : \mu_1 \geq \mu_2$ or $\mu_1 \leq \mu_2$ or $\mu_1 \neq \mu_2$. We studied how to conduct testing when it is assumed that $\sigma_1 = \sigma_2$. Now we will study the case when $\sigma_1 \neq \sigma_2$.

We are given $n_1, \bar{x}_1, s_1, n_2, \bar{x}_2, s_2$, and $\alpha$. The test statistic is given by the formula

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

which under $H_0$, has a $t$-distribution with the number of degrees of freedom $df$ approximated by the largest integer such that

$$df \leq \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}.$$

**Example.** A random sample of 12 male customers of a clothing store showed that they spent, on average, \$89 with a standard deviation of \$27.50. Another random sample of 16 female customers revealed a sample mean of \$105 with a standard deviation of \$36.60. We want to test whether the average amount spent by males is at most that spent by females. We assume that $\sigma_1 \neq \sigma_2$ and $\alpha = 0.05$. Since sample sizes are small, we need to make an additional assumption that the number of hours is normally distributed. We test $H_0: \ \mu_1 = \mu_2$ against $H_1: \ \mu_1 < \mu_2$. We compute the test statistic as

$$t = \frac{89-105}{\sqrt{\frac{(27.50)^2}{12}+\frac{(36.60)^2}{16}}} = -1.3208.$$

The number of degrees of freedom is the largest integer satisfying

$$df \ \leq \ \frac{\left(\frac{(27.50)^2}{12}+\frac{(36.60)^2}{16}\right)^2}{\frac{\left(\frac{(27.50)^2}{12}\right)^2}{12-1}+\frac{\left(\frac{(36.60)^2}{16}\right)^2}{16-1}} \ = \ 25.9957.$$

That is $df = 25$. The $p$-value$= \mathbb{P}(T < -1.3208) = 0.09923 > 0.05$. Thus, we fail to reject $H_0$ and conclude that there is no supporting evidence that the average amount spent by males doesn't exceed that spent by females. $\square$

**Example.** A sample of 50 seniors and a sample of 40 freshmen were surveyed, and it was found that the mean time seniors spend studying for a final exam is 17 hours with a standard deviation of 6 hours, while freshmen spend, on average, 14 hours with a standard deviation of 9 hours. We would like to test whether seniors spend, on average, more hours studying for the final exam than freshmen. We will assume that $\sigma_1 \neq \sigma_2$. We will take $\alpha = 0.05$. We need to test $H_0: \ \mu_1 = \mu_2$ against $H_1: \ \mu_1 > \mu_2$. The test statistic is

$$t = \frac{17-14}{\sqrt{\frac{6^2}{50}+\frac{9^2}{40}}} = 1.811.$$

The number of degrees of freedom is the largest integer such that

$$df \leq \frac{\left(\frac{6^2}{50}+\frac{9^2}{40}\right)^2}{\frac{\left(\frac{6^2}{50}\right)^2}{50-1}+\frac{\left(\frac{9^2}{40}\right)^2}{40-1}} = 65.1,$$

so $df = 65$. The $p$-value$= \mathbb{P}(T > 1.811) < 0.05 = \alpha$ (in Excel, $p$-value$= 0.0374$. We reject the null and conclude that the average number of hours that seniors spend studying for the final exam is larger than that for freshmen. $\square$

**Example.** A study found that the mean number of children under 18 per household in Community A was 1.6 with a standard deviation of 0.7. In Community B, the mean was 2.1 with a standard deviation of 1.3. The data were based on two independent samples of sizes 155 and 160, respectively. We are interested in testing whether the population means differ. There is no reason to believe that the population standard deviations are equal, so we assume $\sigma_1 \neq \sigma_2$ and $\alpha = 0.05$. We write $H_0: \ \mu_1 = \mu_2$ against $H_1: \ \mu_1 \neq \mu_2$. The test statistic is found as

$$t = \frac{1.6 - 2.1}{\sqrt{\frac{0.7^2}{155}+\frac{1.3^2}{160}}} = -4.268.$$

The number of degrees of freedom is the largest integer such that

$$df \leq \frac{\left(\frac{0.7^2}{155}+\frac{1.3^2}{160}\right)^2}{\frac{\left(\frac{0.7^2}{155}\right)^2}{155-1}+\frac{\left(\frac{1.3^2}{160}\right)^2}{160-1}} = 245.7,$$

so $df = 245$. The $p$-value$= 2\mathbb{P}(T < -4.268) < 0.001 < 0.05 = \alpha$ (from Excel, $p$-value$= 0.00003$). We reject $H_0$ and conclude that the population mean number of children is different for these two communities. $\square$

### Confidence Interval for Difference in Means when Standard Deviations are Unequal

**Definition.** A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ when $\sigma_1 \neq \sigma_2$ is

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2,df}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

where the number of degrees of freedom is the largest integer satisfying

$$df \leq \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}.$$

**Example.** From the previous example, we have $df = 245$, and so the critical value is $t_{0.025,245} = 1.9697$. We compute a 95% CI for $\mu_1 - \mu_2$ as

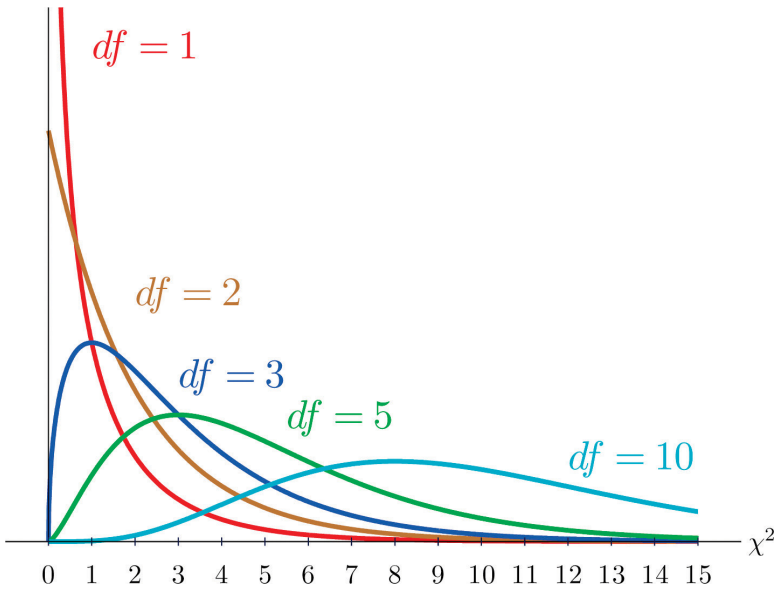
$$1.6 - 2.1 \pm (1.9697)\sqrt{\frac{0.7^2}{155} + \frac{1.3^2}{160}} = -0.5 \pm 0.23 = [-0.73, -0.27].$$

This confidence interval doesn't include 0, and thus we reject the null and draw the same conclusion as in the previous example. □

## THE CHI-SQUARED DISTRIBUTION

**Definition.** A continuous random variable $X$ has a chi-squared distribution with $k$ degrees of freedom (write $X \sim \chi^2(k)$) if the pdf is

$$f_X(x) = \frac{x^{k/2-1}}{\Gamma(k/2)2^{k/2}} e^{-x/2}, \ x > 0.$$




**Note.** Note that a chi-squared distribution with $k$ degrees of freedom is, in fact, a gamma distribution with parameters $\alpha = k/2$ and $\beta = 2$. In addition, the sum of squared of $n$ independent standard normal random variables has a chi-squared distribution with $n$ degrees of freedom, that

is,

$$\sum_{i=1}^{n} Z_i^2 \sim \chi^2(n), \text{ where } Z_i \overset{iid}{\sim} N(0,1),\ i = 1, \ldots, n.$$

**Note.** The chi-square distribution was discovered in 1863 by Ernst Karl Abbe (1840 – 1905) who was a German physicist.

## CONFIDENCE INTERVALS FOR VARIANCE AND STANDARD DEVIATION

**Theorem.** Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ (or $n \geq 30$ and so, by the CLT, the distribution is approximately normal). Then

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

"Proof": We give a proof that is not very rigorous mathematically. It is not difficult to show algebraically that

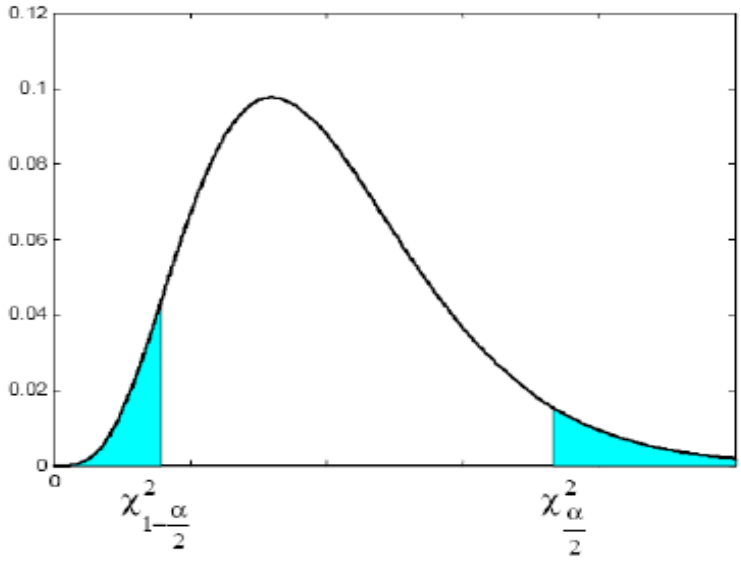
$$\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 - \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2.$$

The first term has a $\chi^2(n)$ distribution, the second term has a $\chi^2(1)$ distribution, and the two terms are independent (the proof of independence is non-trivial). Further, it can be shown (using the moment generating function, for instance) that the difference between $\chi^2(n)$ and $\chi^2(1)$ independent random variables has a $\chi^2(n-1)$ distribution. $\square$

By this theorem, we can use the quantity $(n-1)s^2/\sigma^2$ as a pivot to construct a confidence interval for $\sigma^2$. We write

$$\mathbb{P}\left(\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right) = 1 - \alpha,$$

where $\mathbb{P}(X > \chi^2_{1-\alpha/2}) = 1 - \alpha/2$ and $\mathbb{P}(X > \chi^2_{\alpha/2}) = \alpha/2$, that is, $\chi^2_{1-\alpha/2}$ is the lower critical value and $\chi^2_{\alpha/2}$ is the upper critical value (see the figure).




**Note.** The critical values can be looked up in the chi-squared table. In Excel, we can type "=chisq.inv(prob,df)".

**Definition.** A $100(1-\alpha)\%$ CI for variance $\sigma^2$ is

$$\Big[\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}},\ \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\Big].$$

**Definition.** A $100(1-\alpha)\%$ CI for standard deviation $\sigma$ is

$$\Big[\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}},\ \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}}\Big].$$

**Example.** Suppose we want to compute 95% confidence intervals for the population variance and standard deviation of the volume of 20-ounce Coke bottles. A sample of size 60 bottles was measured and it was obtained that the sample standard deviation is 0.6 oz. We compute$\chi^2_{0.975,59} = 39.6619$ and $\chi^2_{0.025,59} = 82.1174$. Thus, a 95% CI for $\sigma^2$ is

$$\Big[\frac{(60-1)(0.6)^2}{82.1174},\ \frac{(60-1)(0.6)^2}{39.6619}\Big] = [0.26,\ 0.53],$$

and a 95% CI for $\sigma$ is

$$\big[\sqrt{0.26},\ \sqrt{0.53}\big] = [0.51,\ 0.73]. \quad \square$$

## THE CHI-SQUARED TEST FOR VARIANCE

Suppose we would like to test $H_0:\ \sigma^2 = \sigma_0^2$ against $H_1:\ \sigma^2 > \sigma_0^2$ or $\sigma^2 < \sigma_0^2$ or $\sigma^2 \neq \sigma_0^2$. The test statistic is

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

which under $H_0$, has a $\chi^2$ distribution with $n-1$ degrees of freedom. Since the chi-squared distribution is not symmetric, for this test, we can't compute $p$-values. We conduct this test based on the rejection region which is defined as

$$RR = \begin{cases} \{\chi^2 \,|\, \chi^2 > \chi^2_{\alpha,n-1}\}, & \text{if } H_1:\ \sigma^2 > \sigma_0^2, \\ \{\chi^2 \,|\, \chi^2 < \chi^2_{1-\alpha,n-1}\}, & \text{if } H_1:\ \sigma^2 < \sigma_0^2, \\ \{\chi^2 \,|\, \chi^2 < \chi^2_{1-\alpha/2,n-1} \text{ or } \chi^2 > \chi^2_{\alpha/2,n-1}\}, & \text{if } H_1:\ \sigma^2 \neq \sigma_0^2. \end{cases}$$

**Example.** A sample of size $n = 22$ revealed a sample variance of $s^2 = 1.98$. We need to test $H_0: \sigma^2 = 2.3$ against $H_1:\ \sigma^2 < 2.3$. We compute the test statistic $\chi^2 = \dfrac{(22-1)(1.98)}{2.3} = 18.08$. The critical value is $\chi^2_{0.95,21} = 11.59$. The test statistic is larger than the critical value. It means that the test statistic is not in the rejection region, and therefore, we fail to reject the null hypothesis and conclude that the population variance is not less than 2.3. □

**Example.** Suppose that 100 observations produce the sample standard deviation of 6.3. We need to test $H_0:\ \sigma = 5$ against $H_1:\ \sigma > 5$. The test statistic is $\chi^2 = \dfrac{(100-1)(6.3)^2}{5^2} = 157.17$. The rejection region is $RR = \{\chi^2 \,|\, \chi^2 > \chi^2_{0.05,99}\} = \{\chi^2 \,|\, \chi^2 > 123.2252\}$. The observed test statistic is in the rejection region, hence we reject $H_0$, and conclude that the population standard deviation is larger than 5. □

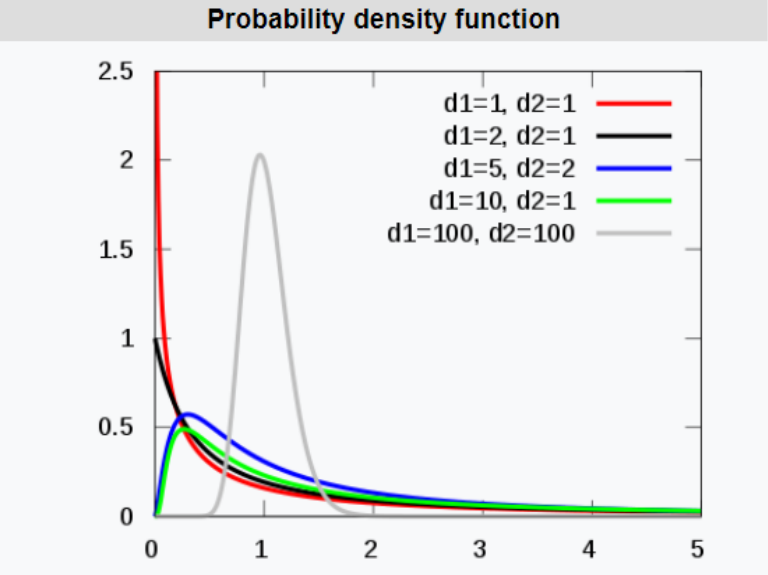
**Example.** In a sample of size 50, the sample variance is 0.78. We need to test $H_0: \sigma^2 = 2$ against $H_1:\ \sigma^2 \neq 2$. We compute the test statistic $\chi^2 = \dfrac{(50-1)(0.78)}{2} = 19.11$. The rejection region is $RR = \{\chi^2 \,|\, \chi^2 < \chi^2_{0.975,49} \text{ or } \chi^2_{0.025,49}\} = \{\chi^2 \,|\, \chi^2 < 31.55 \text{ or } 70.22\}$. The test statistic belongs to the rejection region, therefore, we reject $H_0$ and conclude that the population variance is different from 2. □

## THE F-DISTRIBUTION

**Definition.** The $F$**-distribution** has the probability density function

$$f(x) = \frac{(a/b)^{a/2}}{B(a/2, b/2)}\, x^{a/2-1}(1 + ax/b)^{-(a+b)/2}, \quad a, b, x > 0.$$

Here $B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1}\, dx$ is the **beta function**.




**Note.** The $F$-distribution is named after Sir Ronald Aylmer Fisher (1890 – 1962), a famous British statistician.

**Theorem.** Consider two independent random variables $X_1$ and $X_2$ that have chi-squared distributions with $df_1$ and $df_2$ degrees of freedom, respectively. Then

$$F = \frac{X_1/df_1}{X_2/df_2}$$

has an $F(df_1, df_2)$ distribution. The parameters $df_1$ and $df_2$ are termed the **degrees of freedom of numerator and denominator**, respectively.

## CONFIDENCE INTERVALS FOR RATIO OF VARIANCES AND STANDARD DEVIATIONS

Suppose two independent samples of sizes $n_1$ and $n_2$ are drawn and the sample standard deviations $s_1$ and $s_2$ are measured. We want to construct $100(1-\alpha)\%$ CIs for $\sigma_1^2/\sigma_2^2$ and for $\sigma_1/\sigma_2$. Consider $X_1 = \dfrac{(n_1-1)s_1^2}{\sigma_1^2} \sim \chi^2(n_1-1)$ which is independent of $X_2 = \dfrac{(n_2-1)s_2^2}{\sigma_2^2} \sim \chi^2(n_2-1)$. By the above theorem,

$$F = \frac{(n_1-1)s_1^2}{\sigma_1^2\,(n_1-1)} \div \frac{(n_2-1)s_2^2}{\sigma_2^2\,(n_2-1)} = \frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1, n_2-1).$$

Note that $F$ is a pivotal quantity. We can construct a confidence interval based on $F$. We write

$$\mathbb{P}\Big(F_{1-\alpha/2,n_1-1,n_2-1} < \frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2} < F_{\alpha/2,n_1-1,n_2-1}\Big) = 1-\alpha,$$

$$\mathbb{P}\Big(\frac{s_1^2/s_2^2}{F_{\alpha/2,n_1-1,n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2/s_2^2}{F_{1-\alpha/2,n_1-1,n_2-1}}\Big) = 1-\alpha.$$

**Definition.** A $100(1-\alpha)\%$ confidence interval for $\sigma_1^2/\sigma_2^2$ is

$$\Big[\frac{s_1^2/s_2^2}{F_{\alpha/2,n_1-1,n_2-1}},\ \frac{s_1^2/s_2^2}{F_{1-\alpha/2,n_1-1,n_2-1}}\Big].$$

**Note.** To compute a critical value in Excel, we can enter "=f.inv(prob,df1,df2)". Critical values for $F$-distribution are tabulated for different values of $df_1$ and $df_2$. Occasionally, the following result may be used to extract critical values from the table.

**Result.** The following relation holds for $F$-distribution:

$$F_{\alpha,df_1,df2} = \frac{1}{F_{1-\alpha,df_2,df1}}.$$

This relation can be shown by noting that for $F = \frac{X_1/df_1}{X_2/df_2} \sim F(df_1, df_2)$, $\mathbb{P}(F > F_{\alpha,df_1,df_2}) = \alpha$, and so,

$$\mathbb{P}\Big(\frac{X_1/df_1}{X_2/df_2} > F_{\alpha,df_1,df_2}\Big) = \alpha,$$

from where,

$$\mathbb{P}\Big(\frac{X_2/df_2}{X_1/df_1} < \frac{1}{F_{\alpha,df_1,df_2}}\Big) = \alpha.$$

Thus,

$$\frac{1}{F_{\alpha,df_1,df_2}} = F_{1-\alpha,df_2,df_1}.$$

**Definition.** A $100(1-\alpha)\%$ confidence interval for $\sigma_1/\sigma_2$ is

$$\Big[\sqrt{\frac{s_1^2/s_2^2}{F_{\alpha/2,n_1-1,n_2-1}}},\ \sqrt{\frac{s_1^2/s_2^2}{F_{1-\alpha/2,n_1-1,n_2-1}}}\Big] = \Big[\frac{s_1/s_2}{\sqrt{F_{\alpha/2,n_1-1,n_2-1}}},\ \frac{s_1/s_2}{\sqrt{F_{1-\alpha/2,n_1-1,n_2-1}}}\Big].$$

**Example.** Suppose $n_1 = 65, n_2 = 68, s_1^2 = 37.1$, and $s_2^2 = 42.3$. A 95% CI for $\sigma_1^2/\sigma_2^2$ is

$$\Big[\frac{37.1/42.3}{F_{0.025,64,67}},\ \frac{37.1/42.3}{F_{0.975,64,67}}\Big] = \Big[\frac{37.1/42.3}{1.628},\ \frac{37.1/42.3}{0.6126}\Big] = [0.54, 1.43],$$

and a 95% CI for $\sigma_1/\sigma_2$ is $[\sqrt{0.54}, \sqrt{1.43}] = [0.73, 1.20]$. □

## THE F-TEST FOR RATIO OF TWO VARIANCES

Suppose we need to test $H_0: \sigma_1^2 = \sigma_2^2$ against $H_1: \sigma_1^2 > \sigma_2^2$ or $\sigma_1^2 < \sigma_2^2$ or $\sigma_1^2 \neq \sigma_2^2$. To obtain the test statistic we use the fact that $F = \dfrac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2}$ has an $F$-distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. Assuming $H_0: \sigma_1^2 = \sigma_2^2$ is true, $F$ simplifies to $F = s_1^2/s_2^2$. This is the test statistic. The rejection region for this test is defined as

$$RR = \begin{cases} \{F \,|\, F > F_{\alpha,n_1-1,n_2-1}\}, & \text{if } H_1: \sigma_1^2 > \sigma_2^2, \\ \{F \,|\, F < F_{1-\alpha,n_1-1,n_2-1}\}, & \text{if } H_1: \sigma_1^2 < \sigma_2^2, \\ \{F \,|\, F < F_{1-\alpha/2,n_1-1,n_2-1} \text{ or } F > F_{\alpha/2,n_1-1,n_2-1}\}, & \text{if } H_1: \sigma_1^2 \neq \sigma_2^2. \end{cases}$$

**Example.** Suppose $n_1 = 15, n_2 = 13, s_1 = 4.7$, and $s_2 = 8.9$. We want to test whether $H_0: \sigma_1 = \sigma_2$ vs. $H_1: \sigma_1 < \sigma_2$. The test statistic is $F = (4.7)^2/(8.9)^2 = 0.2789$. The critical value for the rejection region is $F_{0.95,14,12} = 0.3946$. The observed test statistic is below the critical value, therefore we reject the null in favor of the alternative. □

**Example.** Suppose $n_1 = 45, n_2 = 45, s_1^2 = 65.7$, and $s_2^2 = 40.3$. We test $H_0: \sigma_1^2 = \sigma_2^2$ against $H_1: \sigma_1^2 > \sigma_2^2$. The test statistic is $F = 65.7/40.3 = 1.63$. The critical value for the rejection region is $F_{0.05,44,44} = 1.65$. The test statistic is below the critical value, thus we fail to reject the null hypothesis and conclude that variances are equal. □

**Example.** Suppose $n_1 = 8, n_2 = 11, s_1 = 12.3$ and $s_2 = 16.7$. We need to test $H_0: \sigma_1 = \sigma_2$ against $H_1: \sigma_1 \neq \sigma_2$. The test statistic is $F = (12.3)^2/(16.7)^2 = 0.54$. The critical values of the rejection region are $F_{0.975,7,10} = 0.21$ and $F_{0.025,7,10} = 3.95$. The test statistic is not in the rejection region, thus we fail to reject $H_0$ and conclude that population standard deviations don't differ. □

## LIKELIHOOD RATIO TEST

**Definition.** Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$, and let $L(\theta) = \prod_{i=1}^n f(X_i;\theta)$ be the likelihood function. Suppose we want to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. We define the test statistic that is termed the **likelihood ratio (LR)**

$$\Lambda = \frac{L(\theta) \text{ under } H_0}{L(\theta) \text{ under } H_1} = \frac{L(\theta_0)}{L(\widehat{\theta}_{MLE})} = \frac{\prod_{i=1}^n f(X_i;\theta_0)}{\prod_{i=1}^n f(X_i;\widehat{\theta}_{MLE})}.$$

Under $H_0 : \theta = \theta_0$, assuming that $\widehat{\theta}_{MLE}$ is close to $\theta_0$, the likelihood ratio $\Lambda$ should be close to 1. Thus, $H_0$ should be rejected if $\Lambda$ is small. We define the rejection region by $RR = \{\Lambda \,|\, \Lambda \leq c\}$ where the critical value $c$ is found from the expression for the significance level $\alpha = \mathbb{P}(\Lambda \in RR \,|\, H_0 \text{ is true}) = \mathbb{P}(\Lambda \leq c \,|\, \theta = \theta_0)$. The test with this rejection region is called the **likelihood ratio test (LRT)**.

**Note.** The likelihood ratio test is hard to implement in practice because it is difficult to find the exact distribution of $\Lambda$. Instead, an asymptotic likelihood ratio test is implemented.

**Definition.** Suppose we want to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ at a significance level $\alpha$. An **asymptotic likelihood ratio test** has the test statistic $\chi^2 = -2\ln\Lambda$ which, under $H_0$ has a chi-squared distribution with one degree of freedom. The rejection region has the form $RR = \left\{\chi^2 \,|\, \chi^2 > \chi^2_{\alpha,1}\right\}$. For $\alpha = 0.05$, $\chi^2_{0.05,1} = 3.84146$.

**Example (Bernoulli distribution).** Suppose $X_1, \ldots, X_n \overset{iid}{\sim} Ber(p)$ and we want to test $H_0 : p = p_0$ against $H_1 : p \neq p_0$ at the 5% level of significance. First, we compute the likelihood ratio ($\widehat{p}_{MLE} = \bar{X}$)

$$\Lambda = \frac{\prod_{i=1}^n p_0^{X_i}(1-p_0)^{1-X_i}}{\prod_{i=1}^n \bar{X}^{X_i}(1-\bar{X})^{1-X_i}} = \left(\frac{p_0}{\bar{X}}\right)^{n\bar{X}}\left(\frac{1-p_0}{1-\bar{X}}\right)^{n-n\bar{X}}.$$

Then we compute the asymptotic likelihood ratio test statistic $\chi^2 = -2\ln(\Lambda)$, and check if it belongs to the rejection region $RR = \{\chi^2 : \chi^2 > 3.84146\}$.

To consider a numerical example, suppose we flip a coin and observed the sequence ($1 = H, 0 = T$)

$$0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1$$

and would like to test if it is a fair coin, that is, we are interested in testing $H_0 : p = 0.5$ against $H_1 : p \neq 0.5$. We have $n = 14, \bar{X} = 9/14$, and $p_0 = 0.5$. The likelihood ratio is $\Lambda = \left(\frac{0.5}{9/14}\right)^9\left(\frac{1-0.5}{1-9/14}\right)^{14-9} = 0.5602$. The asymptotic likelihood test statistic is $\chi^2 = -2\ln(0.5602) = 1.1589$. It doesn't fall in the rejection region, therefore, we fail to reject the null hypothesis and conclude that the coin is fair. $\square$

**Example (geometric distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} geom(p)$. Suppose we want to test $H_0: p = p_0$ against $H_1: p \neq p_0$. We compute the LR

$$\Lambda = \frac{\prod_{i=1}^n p_0(1-p_0)^{X_i-1}}{\prod_{i=1}^n \left(\frac{1}{\bar{X}}\right)\left(1-\frac{1}{\bar{X}}\right)^{X_i-1}} = \frac{p_0^n(1-p_0)^{n\bar{X}-n}}{\frac{1}{\bar{X}^n}\left(\frac{\bar{X}-1}{\bar{X}}\right)^{n\bar{X}-n}} = p_0^n \bar{X}^{n\bar{X}}\left(\frac{1-p_0}{\bar{X}-1}\right)^{n\bar{X}-n}.$$

For instance, a coin is flipped until a head appears. Suppose the experiment was repeated 25 times and it took 3.2 flips, on average. We want to test if the flipped coin is fair. We are given $n = 25$, and $\bar{X} = 3.2$. We test $H_0: p = 0.5$ vs. $H_1: p \neq 0.5$, and so $p_0 = 0.5$. The LR is $\Lambda = 0.5^{25}\, 3.2^{(25)(3.2)}\left(\frac{1-0.5}{3.2-1}\right)^{(25)(3.2)-25} = 0.0031$. The test statistic for the asymptotic LRT is $\chi^2 = -2\ln(0.0031) = 11.5297 > 3.84146$, so it belongs to the rejection region. Thus, we reject $H_0$ in favor of the alternative, and conclude that the coin is not fair. □

**Example (Poisson distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} Poi(\lambda)$. Suppose we want to test $H_0: \lambda = \lambda_0$ against $H_1: \lambda \neq \lambda_0$. The likelihood ratio is

$$\Lambda = \frac{\prod_{i=1}^n \frac{\lambda_0^{X_i}}{X_i!} e^{-\lambda_0}}{\prod_{i=1}^n \frac{\bar{X}^{X_i}}{X_i!} e^{-\bar{X}}} = \left(\frac{\lambda_0}{\bar{X}}\right)^{n\bar{X}} e^{-n(\lambda_0-\bar{X})}.$$

As a numerical illustration, suppose we observe $0, 4, 2, 1, 2, 0, 1, 0, 3, 2, 1, 0, 0, 2, 1, 1, 5, 0, 1, 1,$ and would like to test $H_0: \lambda = 2$ against $H_1: \lambda \neq 2$. We compute $n = 20, \bar{X} = 27/20 = 1.35$, and $\lambda_0 = 2$. The likelihood ratio is $\Lambda = \left(\frac{2}{1.35}\right)^{27} e^{-(20)(2-1.35)} = 0.0918$. The chi-squared statistics for the asymptotic LRT is $\chi^2 = -2\ln(0.0918) = 4.7757 > 3.84146$, so it lies inside the rejection region, and thus, we reject the null and conclude that the population parameter is not equal to 2. □

**Example (uniform distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} Unif(0, \theta)$. We would like to test $H_0: \theta = \theta_0$ against $H_1: \theta \neq \theta_0$. The LR is

$$\Lambda = \frac{\prod_{i=1}^n \theta_0^{-1}\mathbb{I}(0 \leq X_i \leq \theta_0)}{\prod_{i=1}^n X_{(n)}^{-1}\mathbb{I}(0 \leq X_i \leq X_{(n)})} = \left(\frac{X_{(n)}}{\theta_0}\right)^n \mathbb{I}(X_{(n)} \leq \theta_0).$$

For example, we observed a sample of size 20 and the observed maximum is 7.1. We would like to test $H_0: \theta = 7.6$ against $H_1: \theta \neq 7.6$. The LR is $\Lambda = \left(\frac{7.1}{7.6}\right)^{20} = 0.2564$. The asymptotic LRT statistic is $\chi^2 = -2\ln(0.2564) = 2.7221 < 3.84146$. It means that the test statistic doesn't belong to the rejection region. We fail to reject $H_0$ and conclude that the parameter is equal to 7.6.

Note that if we, for example, wanted to test $H_0: \theta = 7$ against $H_1: \theta \neq 7$, then our LR would be equal to 0, and thus $\chi^2$ would be equal to infinity and we would reject the null in favor of the

alternative. So, if $\theta_0 < X_{(n)}$, we necessarily conclude that $\theta \neq \theta_0$, which makes sense. $\square$

**Example (exponential distribution).** Let $X_1, \ldots, X_n \overset{iid}{\sim} Exp(\text{mean}=\beta)$. And suppose we want to test $H_0: \beta = \beta_0$ against $H_1: \beta \neq \beta_0$. We compute the LR as

$$\Lambda = \frac{\prod_{i=1}^n \beta_0^{-1} \exp\{-X_i/\beta_0\}}{\prod_{i=1}^n \bar{X}^{-1} \exp\{-X_i/\bar{X}\}} = \Big(\frac{\bar{X}}{\beta_0}\Big)^n \exp\{-n\bar{X}/\beta_0 + n\bar{X}/\bar{X}\} = \Big(\frac{\bar{X}}{\beta_0}\Big)^n \exp\{-n\bar{X}/\beta_0 + n\}.$$

For example, if in a sample of size 10, the mean wait time is 2.2 minutes. We would like to test $H_0: \beta = 3.5$ against $H_1: \beta \neq 3.5$. The LR is $\Lambda = \Big(\frac{2.2}{3.5}\Big)^{10} \exp\{-(10)(2.2)/3.5 + 10\} = 0.3950$. The asymptotic LRT test statistic is $\chi^2 = -2\ln(\Lambda) = -2\ln(0.3950) = 1.8575 < 3.84146$. Thus we don't reject $H_0$ and conclude that the mean is 3.5. $\square$

**Example (normal distribution).** Consider $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ where $\sigma^2$ is known. Suppose we want to test $H_0: \mu = \mu_0$ against $H_1: \mu \neq \mu_0$ at the significance level $\alpha$. One way to test it is to conduct a one-sample z-test with the test statistic $z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ and the rejection region $RR = \{z : |z < z_{1-\alpha/2} \text{ or } z > z_{\alpha/2}\} = \{z : | |z| > z_{\alpha/2}\}$. We will now show that this test is equivalent to the asymptotic likelihood ratio test. We obtain the likelihood ratio as

$$\Lambda = \frac{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{-\frac{(X_i-\mu_0)^2}{2\sigma^2}\Big\}}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{-\frac{(X_i-\bar{X})^2}{2\sigma^2}\Big\}} = \exp\Big\{-\sum_{i=1}^n \frac{(X_i-\mu_0)^2}{2\sigma^2} + \sum_{i=1}^n \frac{(X_i-\bar{X})^2}{2\sigma^2}\Big\}$$

$$= \exp\Big\{-\frac{1}{2\sigma^2}\Big[\sum_{i=1}^n X_i^2 - 2\mu_0 \sum_{i=1}^n X_i + n\mu_0^2 - \sum_{i=1}^n X_i^2 + 2\bar{X}\sum_{i=1}^n X_i - n\bar{X}^2\Big]\Big\}$$

$$= \exp\Big\{-\frac{1}{2\sigma^2}\Big[-2\mu_0\, n\bar{X} + n\mu_0^2 + 2\bar{X}\, n\bar{X} - n\bar{X}^2\Big]\Big\}$$

$$= \exp\Big\{-\frac{n}{2\sigma^2}\Big[\bar{X}^2 - 2\mu_0\,\bar{X} + \mu_0^2\Big]\Big\} = \exp\Big\{-\frac{n}{2\sigma^2}\big(\bar{X} - \mu_0\big)^2\Big\}.$$

The test statistic for the asymptotic likelihood ratio test is

$$\chi^2 = -2\ln\Lambda = \frac{\big(\bar{X} - \mu_0\big)^2}{\sigma^2/n}.$$

This test statistic has an exact $\chi^2(1)$ distribution for any $n$ (not just an asymptotic distribution). The rejection region $RR = \{\chi^2 \,|\, \chi^2 > \chi^2_{\alpha,1}\} = \{z : | |z| > z_{\alpha/2}\}$ because $z^2_{\alpha/2} = \chi_{\alpha,1}$. $\square$

## CHI-SQUARED TESTS FOR CATEGORICAL VARIABLES

**Definition.** A **contingency table** (or a **two-way table**, or a **cross tab**) is a table of frequencies for the level-level combinations of two variables.

A contingency table, its **marginal totals** (row and column totals), and a **grand** (overall) total look like this:

| | | VARIABLE 2 | | | |
|---|---|---|---|---|---|
| | | Level 1 | Level 2 | Level 3 | Total |
| VARIABLE 1 | Level 1 | $n_{11}$ | $n_{12}$ | $n_{13}$ | $n_{1.}$ |
| | Level 2 | $n_{21}$ | $n_{22}$ | $n_{23}$ | $n_{2.}$ |
| | Total | $n_{.1}$ | $n_{.2}$ | $n_{.3}$ | $n_{..}$ |

There are two types of situations when those contingency tables arise: (1) when one sample is drawn and two variables are measured for each individual, or (2) several independent samples are drawn and a single variable is measured on each individual. In case (1), we would be interested in testing $H_0$ : the two variables are independent against $H_1$ : the two variables are not independent (chi-squared test for independence). In case (2), we would be interested in testing $H_0$ : at each variable level, proportions are the same across the samples against $H_1$ : at each variable level, proportions are not all the same across the samples (chi-squared test for equality of proportions).

**Note.** In the test for independence, the testing is done assuming the independence of variables (i.e., the product of marginal probabilities is equal to the joint probability). In the test for equality of proportions, under the null, the underlying distribution is hypergeometric. Even though the theories behind these two tests are completely different, the test statistics are the same and so the two tests are conducted the same way.

### How the chi-square tests are conducted

Step 1. Arrange the counts (frequencies) in a contingency table. Denote by $n_{ij}$ the observed counts in row $i$ and column $j$.
Step 2. Compute row and column totals, and the grand total. Denote these totals, respectively, $n_{i.}, n_{.j}$, and $n_{..}$.
Step 3. Compute **expected values** for each cell using the formula $e_{ij} = \dfrac{n_{i.}n_{.j}}{n_{..}}$ (as the product of respective row and column totals divided by the grand total). Note that by definition, the expected values must sum up to respective row and column totals.
Step 4. Compute the test statistic

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - e_{ij})^2}{e_{ij}},$$

which under $H_0$ has a $\chi^2$-distribution with $df = (r-1)(c-1)$ where $r$ is the number of rows and $c$ is the number of columns.
Step 5. Compute the critical value for the rejection region of the form $RR = \{\chi^2 : \mid \chi^2 > \chi^2_{\alpha,df}\}$. State decision, draw conclusion.

**Note.** The chi-squared test is valid only if the expected counts in each cell are at least 5.

**Example (test for independence).** Suppose a sample of size 65 is chosen and gender (Male/Female) and opinion (Yes/No/Neutral) are collected for each individual in the sample. The data are summarized in the table.

| | Opinion | | | |
|---|---|---|---|---|
| Gender | Yes | No | Neutral | Total |
| Male | 10 | 15 | 17 | 42 |
| Female | 13 | 7 | 3 | 23 |
| Total | 23 | 22 | 20 | 65 |

We want to test $H_0$ : gender and opinion are independent vs. $H_1$ : they are dependent. We compute expected counts and place them in each cell in parentheses like this:

| | Opinion | | | |
|---|---|---|---|---|
| Gender | Yes | No | Neutral | Total |
| Male | 10 (14.86) | 15 (14.22) | 17 (12.92) | 42 |
| Female | 13 (8.14) | 7 (7.78) | 3 (7.08) | 23 |
| Total | 23 | 22 | 20 | 65 |

For instance $e_{11} = (42)(23)/65 = 14.86, e_{12} = (42)(22)/65 = 14.22$, etc. Note that even though one of the observed counts is less than 5, the expected count for that cell is larger than 5, so the chi-squared test is applicable. The test statistic is

$$\chi^2 = \frac{(10-14.86)^2}{14.86} + \frac{(15-14.22)^2}{14.22} + \frac{(17-12.92)^2}{12.92} + \frac{(13-8.14)^2}{8.14} + \frac{(7-7.78)^2}{7.78} + \frac{(3-7.08)^2}{7.08} = 8.25.$$

The number of degrees of freedom is $df = (2-1)(3-1) = 2$. The critical value of the rejection region is $\chi^2_{0.05,2} = 5.99$. Since the observed test statistic falls in the rejection region, we reject the null in favor of the alternative and conclude that gender and opinion are not independent. □

**Example (test for equality of proportions).** Suppose 200 residents of each of the three cities (LA, NY, and Denver) are surveyed and their opinions (Yes/No/No opinion) are recorded. The frequencies are as follows:

| City | Opinion | | | Total |
|---|---|---|---|---|
| | Yes | No | No Opinion | |
| LA | 109 | 82 | 9 | 200 |
| NY | 150 | 35 | 15 | 200 |
| Denver | 122 | 63 | 15 | 200 |
| Total | 381 | 180 | 39 | 600 |

We test $H_0$ : proportions in each column are the same against $H_1$ : in each column, not all proportions are the same. (Note that two proportions can be equal but not equal to the third). We compute the expected counts and add them to the table.

| City | Opinion | | | Total |
|---|---|---|---|---|
| | Yes | No | No Opinion | |
| LA | 109 (127) | 82 (60) | 9 (13) | 200 |
| NY | 150 (127) | 35 (60) | 15 (13) | 200 |
| Denver | 122 (127) | 63 (60) | 15 (13) | 200 |
| Total | 381 | 180 | 39 | 600 |

The expected values are computed as: $(200)(381)/600 = 127$, $(200)(180)/600 = 60$, and $(200)(39)/600 = 13$. The test statistic is

$$\chi^2 = \frac{(109-127)^2}{127} + \frac{(82-60)^2}{60} + \cdots + \frac{(15-13)^2}{13} = 27.4.$$

The number of degrees of freedom is $df = (3-1)(3-1) = 4$. The critical value for the rejection region is $\chi^2_{0.05,4} = 9.48$. Therefore, the observed test statistic belongs to the rejection region, and we reject $H_0$ and conclude that the column proportions are not all equal. □

## CHI-SQUARED GOODNESS-OF-FIT TEST

Suppose we observe some realizations of a random variable and would like to test if that variable follows a specified distribution. This test is called the **goodness-of-fit test** because it tests if the distribution fits the data well. We test $H_0$ the underlying distribution is the specified one against $H_1$ : the underlying distribution is not the specified one. The testing is done based on a chi-squared test statistic $\sum (obs - exp)^2/exp$, which under $H_0$ has a chi-squared distribution with the number of degrees of freedom $df$ =# of categories - 1-# distribution parameters that have to be estimated from the data. The rejection region is of the form $RR = \{\chi^2 \mid \chi^2 > \chi^2_{\alpha,df}\}$.

**Example.** We roll a die 120 times record our observations. We want to test $H_0$ : the die is fair against $H_1$ : the die is not fair. The data are

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Observed counts | 12 | 22 | 8 | 43 | 15 | 20 |

Under $H_0$, each number appears an equal number of times, that is the expected counts all equal to 120/6=20. We have

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Observed counts | 12 | 22 | 8 | 43 | 15 | 20 |
| Expected counts | 20 | 20 | 20 | 20 | 20 | 20 |

The test statistic is

$$\chi^2 = \frac{(12-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(8-20)^2}{20} + \frac{(43-20)^2}{20} + \frac{(15-20)^2}{20} + \frac{(20-20)^2}{20} = 38.3.$$

The number of degrees of freedom $df = 6 - 1 - 0 = 5$. The critical value of the rejection region is $\chi^2_{0.05,5} = 11.07$. The test statistic is in the rejection region and therefore, we reject $H_0$ and conclude that the die is not fair. □

**Example.** Suppose we observe 0,1,1,2,0,3,2,0,1,0,1,2,4,0,5,2,1,1,3,4, and would like to test if these data come from a Poisson distribution. We test $H_0$ : distribution is Poisson against $H_1$ : distribution is not Poisson. We compute the observed frequencies:

| | 0 | 1 | 2 | 3 | 4 | $\geq 5$ |
|---|---|---|---|---|---|---|
| Observed counts | 5 | 6 | 4 | 2 | 2 | 1 |

First we need to estimate $\lambda$ from the data. We have $\widehat{\lambda} = \bar{X} = 33/20 = 1.65$. Then we compute the expected counts based on $Poi(1.65)$ distribution. We obtain

$$(20)p(0) = (20)e^{-1.65} = 3.84, \quad (20)p(1) = (20)(1.65e^{-1.65}) = 6.34, \quad (20)p(2) = (20)\big(\frac{1.65^2}{2}e^{-1.65}\big) = 5.23,$$

$$(20)p(3) = (20)\big(\frac{1.65^3}{6}e^{-1.65}\big) = 2.88, \quad (20)p(4) = (20)\big(\frac{1.65^4}{24}e^{-1.65}\big) = 1.19,$$

and, finally, the last expected count (for 5 or above) can be found by subtraction from 20: $20 - 3.84 - \cdots - 1.19 = 0.52$. Putting the observed and expected counts together in a table, we get

| | 0 | 1 | 2 | 3 | 4 | $\geq 5$ |
|---|---|---|---|---|---|---|
| Observed counts | 5 | 6 | 4 | 2 | 2 | 1 |
| Expected counts | 3.84 | 6.34 | 5.23 | 2.88 | 1.19 | 0.52 |

The test statistic is

$$\chi^2 = \frac{(5-3.84)^2}{3.84} + \frac{(6-6.34)^2}{6.34} + \cdots + \frac{(1-0.52)^2}{0.52} = 1.92.$$

The number of degrees of freedom is $df = 6 - 1 - 1 = 4$. The critical value for the rejection region is $\chi^2_{0.05,4} = 9.49$, thus the observed test statistic doesn't belong to the rejection region, and we fail to reject the null and conclude that the underlying distribution is Poisson. $\square$

**Example.** Suppose we observed the following data

| | | | | |
|---|---|---|---|---|
| 0.17 | 0.33 | 0.54 | 0.59 | 0.63 |
| 0.70 | 0.71 | 0.74 | 0.86 | 0.99 |
| 1.15 | 1.23 | 1.25 | 1.43 | 1.79 |
| 1.82 | 1.83 | 1.86 | 1.89 | 1.90 |

We want to test $H_0: \ X \sim Unif(0,2)$ against $H_1: \ X \not\sim Unif(0,2)$. We divide the interval into subintervals (bins) of equal lengths and compute observed frequencies for each subinterval. We obtain

| | [0,0.4) | [0.4,0.8) | [0.8, 1.2) | [1.2,1.6) | [1.6, 2) |
|---|---|---|---|---|---|
| observed counts | 2 | 6 | 3 | 3 | 6 |
| expected counts | 4 | 4 | 4 | 4 | 4 |

The chi-squared statistics is

$$\chi^2 = \frac{(2-4)^2}{4} + \cdots + \frac{(6-4)^2}{4} = 3.5$$

which under $H_0$ has a chi-squared distribution with $df = 5 - 1 - 0 = 4$. The critical value for the rejection region is $\chi^2_{0.05,4} = 9.49$. The observed test statistic doesn't fall in the rejection region, and thus we fail to reject the null and conclude that the data come from a uniform distribution on (0,2). $\square$

**Example.** Consider the same data as in the previous example. We would like to test $H_0$ : data are exponentially distributed against $H_1$ : data are not exponentially distributed. To fit an exponential distribution, we need to estimate the mean. We write $\widehat{\beta} = \bar{X} = 1.1205$. Next, we compute the expected counts for each bin based on the exponential distribution with a mean of 1.1205.

$$(20)\mathbb{P}(0 < X < 0.4) = (20)\big(1 - e^{-0.4/1.1205}\big) = 6.004,$$

$$(20)\mathbb{P}(0.4 < X < 0.8) = (20)\big(e^{-0.4/1.1205} - e^{-0.8/1.1205}\big) = 4.202,$$

$$(20)\mathbb{P}(0.8 < X1.2) = (20)\big(e^{-0.8/1.1205} - e^{-1.2/1.1205}\big) = 2.940,$$

$$(20)\mathbb{P}(1.2 < X < 1.6) = (20)\big(e^{-1.2/1.1205} - e^{-1.6/1.1205}\big) = 2.058,$$

and

$$(20)\mathbb{P}(X > 1.6) = (20)e^{-1.6/1.1205} = 4.796.$$

Summing it all in a table, we get

| | [0,0.4) | [0.4,0.8) | [0.8, 1.2) | [1.2,1.6) | [1.6, $\infty$) |
|---|---|---|---|---|---|
| observed counts | 2 | 6 | 3 | 3 | 6 |
| expected counts | 6.004 | 4.202 | 2.940 | 2.058 | 4.796 |

The test statistic is

$$\chi^2 = \frac{(2-6.004)^2}{6.004} + \cdots + \frac{(6-4.796)^2}{4.796} = 4.18.$$

The number of degrees of freedom is $df = 5 - 1 - 1 = 3$. The critical value for the rejection region is $\chi^2_{0.05,3} = 7.81$. The observed test statistic doesn't belong to the rejection region, thus we fail to reject the null and conclude that the data are exponentially distributed. $\square$

**THE END**