CS 660: Mathematical Foundations of Analytics

Dr. Francis Parisi

Pace University

Spring 2018

IMS Chapter 4 – Statistical Inference

Statistical Inference

Sampling and Statistics

- In this chapter we will look at some of the tools used in statistical inference
- Typically in a statistical problem we have som erandom variable X but don't know anything about its pmf or pdf
- Basically there's two degrees of ignorance
  1. We don't know anything about $f(x)$ or $p(x)$
  2. We know the form of $f(x)$ or $p(x)$ but don't know some parameter say θ

Statistical Inference

Focusing on the second scenario above, we have some examples

(a) X has an exponential distribution, $Exp(\theta)$, where θ is unknown

(b) X has a binomial distribution $b(n, p)$, where n is known but p is unknown

(c) X has a gamma distribution $\Gamma(\alpha, \beta)$, where both α and β are unknown

(d) X has a normal distribution $N(\mu^2, \sigma^2)$, where both the mean μ and the variance ^{2} of X are unknown

Statistical Inference

We address this problem by saying X has a pdf or pmf of the form $f(x;\theta)$ or $p(x;\theta)$

- $\theta \in \Omega$ for some set Ω
- In this example, $\Omega = \{\theta \,|\, \theta > 0\}$
- θ is the **parameter** of the distribution
- Because θ is unknown, we use statistical inference to estimate it

Statistical Inference

- What we learn about the distribution of X and the parameters of the distribution comes from a **sample** on X
- The observations from the have the same distribution as X, denoted as the random variables $X_1, X_2, \ldots, X_n$, where n denotes the **sample size**
- We use lower case letters $x_1, x_2, \ldots, x_n$ to denote the values or **realizations** of the sample
- We often assume that the sample observations $X_1, X_2, \ldots, Xn$ are also mutually independent, so we call the sample a **random sample**

Statistical Inference

Definition (4.1.1)

If the random variables $X_1, X_2, \ldots, X_n$ are independent and identically distributed (*iid*), then these random variables constitute a **random sample** of size n from the common distribution.

Definition (4.1.2)

Let $X_1, X_2, \ldots, X_n$ denote a sample on a random variable X. Let $T = T(X_1, X_2, \ldots, X_n)$ be a function of the sample. Then T is called a **statistic**.

Statistical Inference

Using the above terminology our problem becomes:

Let $X_1, X_2, \ldots, X_n$ denote a *random sample* on a random variable X with a density or mass function of the form $f(x;\theta)$ or $p(x;\theta)$, where $\theta \in \Omega$ for a specified set Ω.

It makes sense to consider a *statistic* T , which is an **estimator** of θ. More formally, T is called a **point estimator** of θ. While we call T an estimator of θ, we call its realization t an **estimate** of θ.

Statistical Inference

Definition

Let $X_1, X_2, \ldots, X_n$ denote a sample on a random variable X with pdf $f(x;\theta)$, $\theta \in \Omega$. Let $T = T(X_1, X_2, \ldots, X_n)$ be a statistic. We say that T is an unbiased estimator of θ if $\mathbb{E}(T) = \theta$.

Statistical Inference

Estimating θ

- We'll look at the **maximum likelihood estimator (mle)** to find point estimators
- To proceed we consider the joint distribution of our sample; since the observations are *iid* the joint distribution function is the product of the pdfs $\prod_{i=1}^{n} f(x_i; \theta)$
- Then as a function of θ we have

$$L(\theta) = L(\theta; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta)$$

- This is the **likelihood function** of the random sample

Statistical Inference

- We often seek as an estimate the value of θ that maximizes the value of $L(\theta)$
- If we find a unique estimate we call it the **maximum likelihood estimator** (mle) and denote it $\hat{\theta}$ or

$$\hat{\theta} = \text{Argmax}\, L(\theta)$$

Statistical Inference

- Since the likelihood function is a product it is easier to work with the [natural] log of the likelihood, $l(\theta) = \log L(\theta)$
- For most of the models we'll discuss, the pdf (or pmf) is a differentiable function of θ, and frequently $\hat{\theta}$solves the equation

$$\frac{\partial l(\theta)}{\partial \theta} = 0$$

- For vector valued θ the results is a system of equations referred to as the **estimating equations**

Statistical Inference

Example

Suppose the common pdf of the random sample $X_1, X_2, \ldots, X_n$ is the $\Gamma(1, \theta)$ density $f(x) = \theta^{-1} \exp -x/\theta$ with support $0 < x < \infty$. This is the exponential distribution. The log of the likelihood function is given by

$$l(\theta) = \log \prod_{i=1}^{n} \frac{1}{\theta} e^{-x_i/\theta} = -n \log \theta - \theta^{-1} \sum_{i=1}^{n} x_i$$

Statistical Inference

Example (continued)

The partial with respect to θ is

$$\frac{\partial l(\theta)}{\partial \theta} = -n\theta^{-1} + \theta^{-2} \sum_{i=1}^{n} x_i$$

Set this equal to zero and solve for θ to get the statistic $\hat{\theta} = \overline{X}$ as the mle of θ

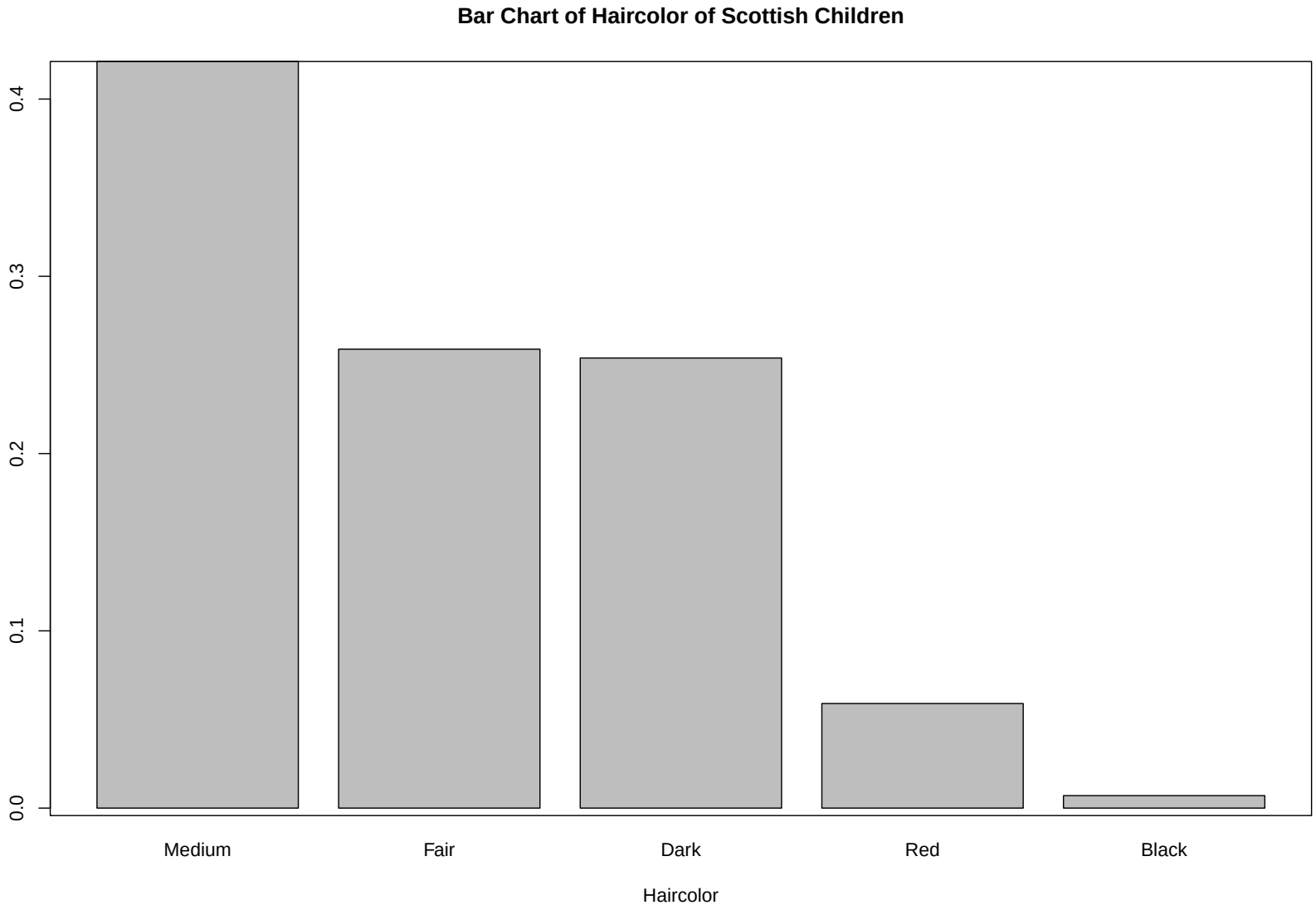
Now $\mathbb{E}[X] = \theta$ and $\mathbb{E}[\overline{X}] = \theta$ so $\hat{\theta}$ is an unbiased estimator of θ

Statistical Inference

- We can visually estimate the pmf or pdf of a random variable using a histogram which is a **nonparametric** estimator
- In the discrete case we can consider the frequency with which the observations fall into specific classes
- In the continuous case we count the observations that fall within specific ranges
- For the coninuous case we use a **kernel density** to smooth the empirical distribution

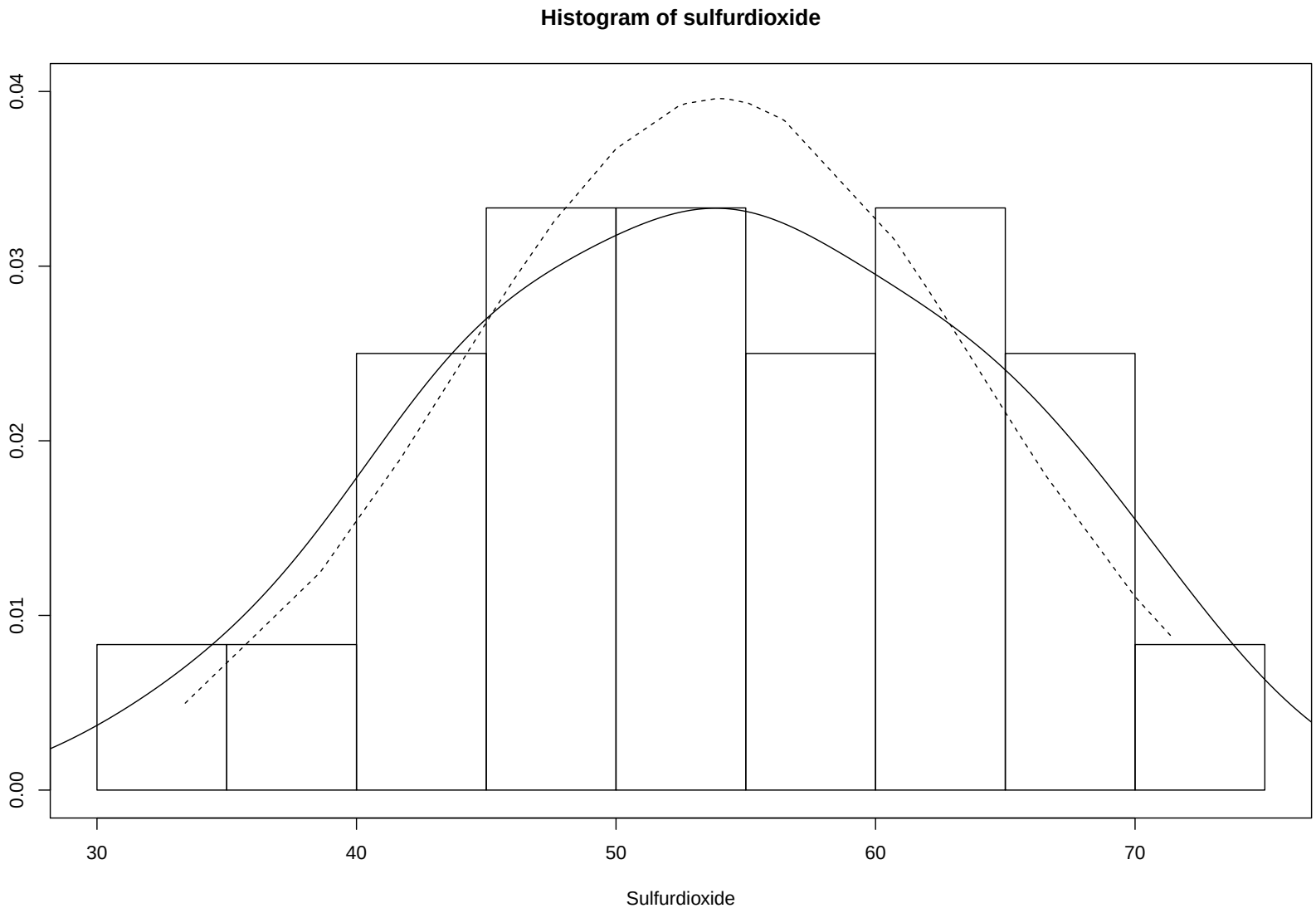
Statistical Inference

Figure 4.1.1: Barchart for discrete X



Statistical Inference

Figure 4.1.2: Histogram of Sulfur Dioxide concentrations with density overlaid (solid line) and a normal density (dashed)



Statistical Inference

Confidence Intervals

- When we estimate a parameter from data we make a point estimate
- But how can we decide just how good this estimate is of the population parameter in which we're interested?
- We can assess the error of this estimate in terms of a confidence interval

Statistical Inference

Definition

Let $X_1, X_2, \ldots, X_n$ be a sample on a random variable X, where X has pdf $f(x;\theta)$, $\theta \in \Omega$. Let $0 < \alpha < 1$ be specified. Let $L = L(X_1, X_2, \ldots, X_n)$ and $U = U(X_1, X_2, \ldots, X_n)$ be two statistics. We say that the interval (L, U) is a $(1-\alpha)100\%$ **confidence interval** for θ if

$$1 - \alpha = P_\theta[\theta \in (L, U)].$$

That is, the probability that the interval includes θ is $1 - \alpha$, which is called the **confidence coefficient** or **confidence level** the of the interval.

Statistical Inference

Another way to interpret the confidence interval is that we repeatedly draw samples and make confidence intervals, say M times, we would expect $(1-\alpha)M$ of these intervals to contain the true value of θ

That is we are $(1-\alpha)100\%$ confident that the true value of θ lies in the interval (l, u)

Statistical Inference

Example (4.2.1)

Suppose the random variables $X_1, X_2, \ldots, X_n$ are a random sample from a $N(\mu, \sigma^2)$ distribution

Let $\overline{X}$ and S^2 denote the sample mean and sample variance, respectively and $\overline{X}$ is the mle of μ and $[(n-1)/n]S^2$ is the mle of σ^2

The random variable $T = (\overline{X} - \mu)/(S/\sqrt{n})$ has a t-distribution with $n-1$ degrees of freedom

Statistical Inference

Example (4.2.1 Continued)

For $0 < \alpha < 1$, define $t_{\alpha/2,n-1}$ to be the upper $\alpha/2$ critical point of a t-distribution with $n-1$ degrees of freedom; i.e., $\alpha/2 = P\left[T > t_{\alpha/2,n-1}\right]$

From this we get

$$
\begin{aligned}
1-\alpha &= P\left[-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}\right] \\
&= P_\mu\left[-t_{\alpha/2,n-1} < \frac{\overline{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2,n-1}\right] \\
&= P_\mu\left[-t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} < \overline{X}-\mu < t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right] \\
&= P_\mu\left[\overline{X}-t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} < \mu < \overline{X}+t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right]
\end{aligned}
$$

Example (4.2.1 Continued)

Once the sample is drawn, let $\bar{x}$ and s denote the realized values of the statistics $\overline{X}$ and S, respectively. Then a $(1-\alpha)100\%$ confidence interval for μ is given by

$$\left(\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right)$$

Statistical Inference

Hypothesis Testing

- Hypothesis testing is a frequently used method of inference
- We start by stating the **null hypothesis**, H_0, and the **alternative hypothesis**, H_1 (sometimes denoted by H_a)
- For example $H_0 : \mu = 0$, and $H_1 : \mu \neq 0$ this is a two-sided ot two-tailed test
- Or $H_1 : \mu < 0$ (one-sided or one-tailed test, left tail), $H_1 : \mu > 0$ (right tail test)
- There are several steps to complete the hypothesis test

Statistical Inference

Steps in hypothesis testing:

Step 1: State the null and alternative hypotheses, H_0, and H_1, determine if this is a one-tailed or two-tailed test

Step 2: State α, the level of significance

Step 3: Calculate the test statistic from the sample mean and variance

$$t = \frac{\overline{X} - \mu}{\sqrt{s^2/n}}$$

Step 4: Compare the calculated test statistic to the critical value

Step 5: Accept H_0 if t is in the acceptance region, or reject H_0 if t is in the critical or rejection region

Statistical Inference

- We set the level of significance α *before* we do the test
- Often $\alpha = 0.05$ and represents the **size** of the critical region
- If the test statistics t falls in the critical region, we reject H_0, otherwise we do not reject H_0
- There's always a risk of making a wrong decision and these are known as **Type I** and **Type II** errors

Statistical Inference

	Reality	
Decision	H_0 is True	H_0 is NOT True
Reject H_0	**Type I** Error	Correct Decision
DO NOT Reject H_0	Correct Decision	**Type II** Error

Statistical Inference

Example

Suppose we sample 26 water bottles for some known chemical. The mean level of the chemical is believed to be 0 mg/dl. We measure each sample and find a sample mean $\overline{X}$ of 1.9 mg/dl of the chemical, with a standard deviation, S equal to 4.2122. Is there sufficient evidence at a $\alpha = 0.05$ level of significance to say the average amount of chemical in the water is not zero?

Statistical Inference

We have $H_0 : \mu = 0$, $H_1 : \mu \neq 0$, $\alpha = 0.05$, $\overline{X} = 1.9$, $S = 4.2122$, and $n = 26$

therefore,

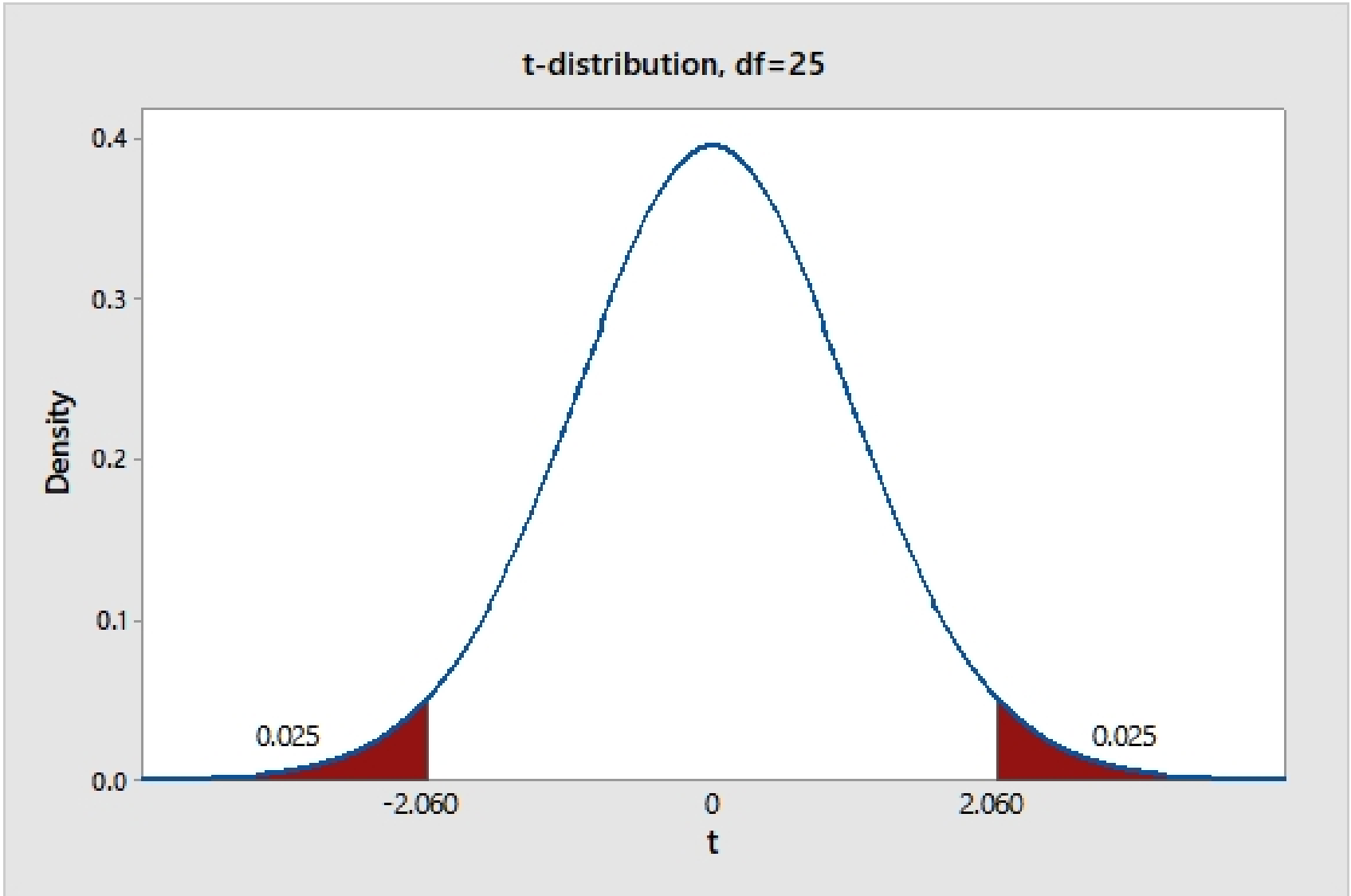
$$t = \frac{1.9 - 0}{4.2122/\sqrt{26}} = 2.3$$

We compare this t to the critical value for $t_{\alpha/2,n-1}$

We find that $t_{\alpha/2,n-1} = 2.06$, thus we reject H_0 since $t > t_{critical}$

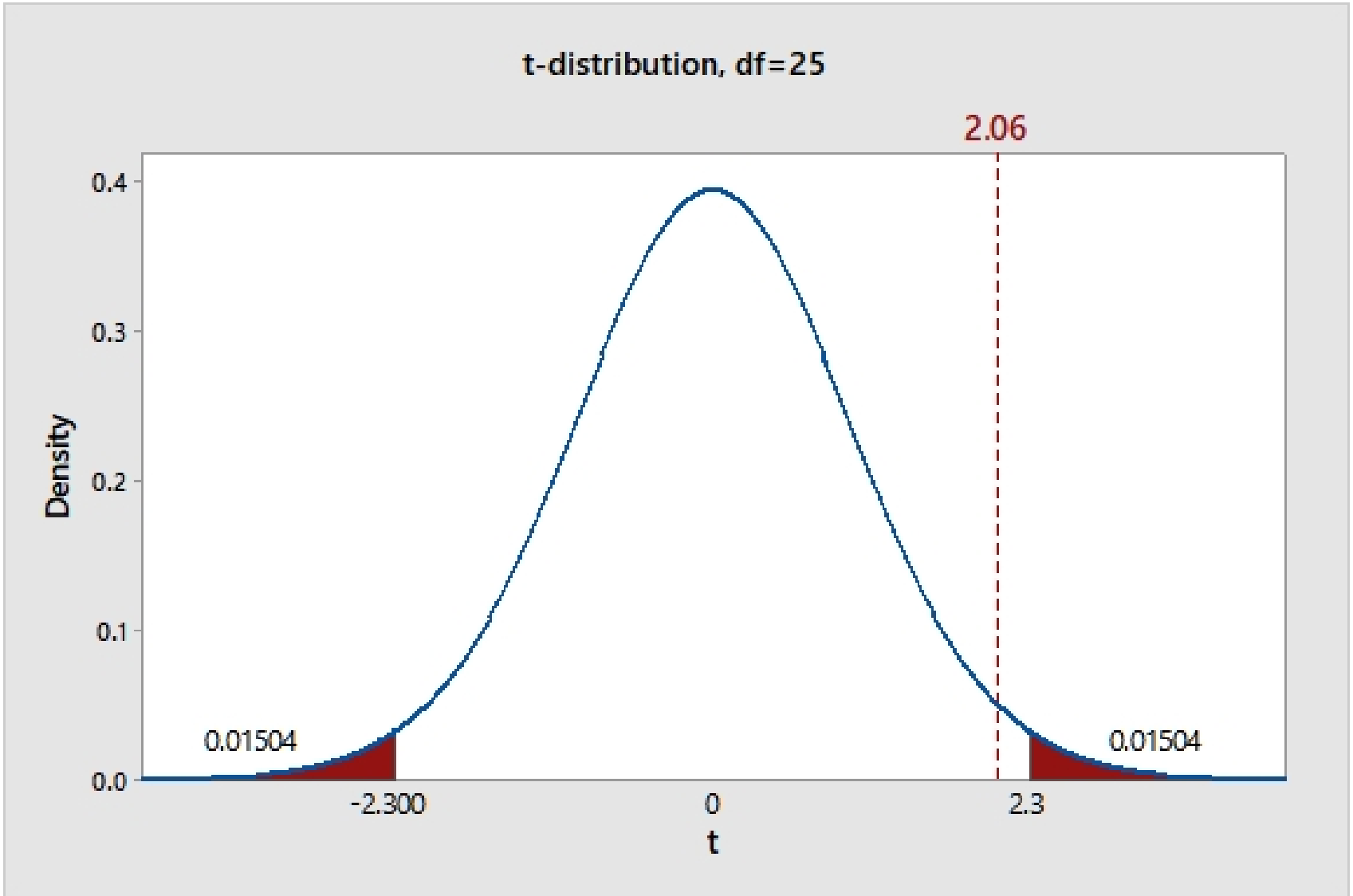
Statistical Inference

What does the critical value mean?



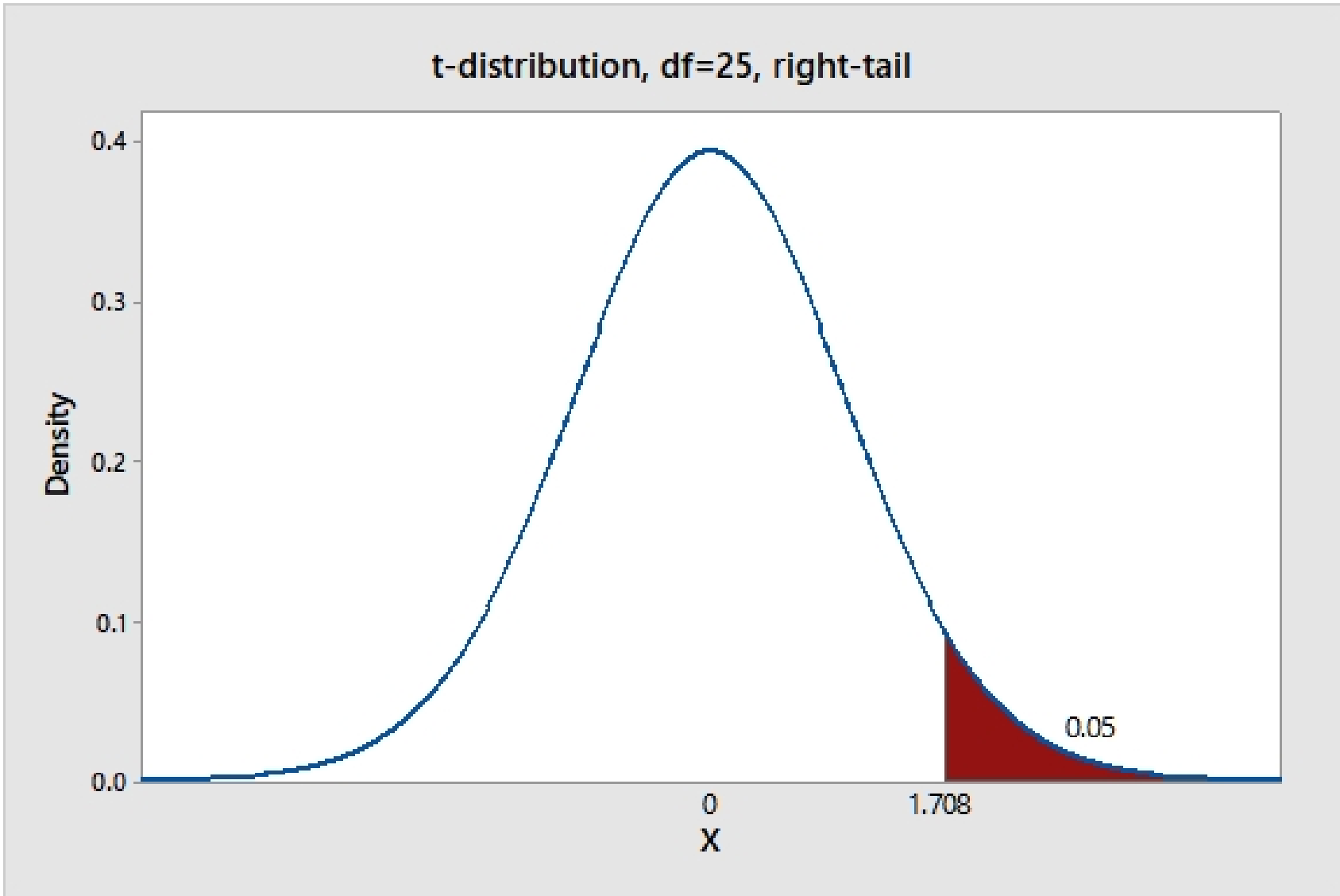
Statistical Inference

p-values



Statistical Inference

For a one-tailed test we have



Statistical Inference

p-values

- The tail probability associated with the t-statistic from our hypothesis test is called the p-value
- If we consider α as our theoretical level of significance then the p-value is the *observed* level of significance
- For a one-tailed test the p-value is the tail probability
- For a two-tailed test it is twice the tail probability
- We *reject the null hypothesis* when $p < \alpha$

References

Hogg, R. V., McKean, J. W. and Craig, A. T. (2018).
Introduction to Mathematical Statistics.
Pearson, Boston, eighth edition.

Ross, S. (2010).
A First Course in Probability.
Prentice Hall, Upper Saddle River, eigth edition.

Wu, J. and Coggeshall, S. (2012).
Foundations of Predictive Analytics.
CRC Press, Boca Raton, FL.