Mathematics/Statistics Bootcamp

Part V: Inference

Steven Winter Christine Shen

Department of Statistical Science
Duke University

MSS Orientation, August 2022

What is statistical inference?

- A **statistical experiment** generates a collection of data $\mathbf{X}$.
- The set of all possible data values is the **sample space** Ω.
- A **statistical model** is a family of possible distributions $\{P_\theta, \theta \in \Theta\}$[1] for $\mathbf{X}$, where Θ denotes the parameter space.

E.g., consider an experiment of tossing a coin n times, and recording *Head* (H) or *Tail* (T) for each toss.

- The sample space is $\Omega = \{H, T\}^n$.
- Assume the tosses are independent with an equal head probability θ. Let $n_h = \sum_{i=1}^n \mathbb{1}(x_i = H)$ denote the total number of heads,

$$P_\theta(X_1 = x_1, \ldots, X_n = x_n) = \theta^{n_h}(1-\theta)^{n-n_h}.$$

[1] P_θ: frequentist notation which refers to the model with parameter value θ

What is Statistical Inference?

We are usually interested in:

- learning about θ, or some function $g(\theta)$ based on the data. Common types of inference problems are:
  1. Point estimation;
  2. Interval estimation;
  3. Hypothesis testing;
  4. Prediction.
- evaluating performance of the inference procedure for
  1. finite sample;
  2. asymptotics (i.e., as number of data n goes to ∞).

Overview

Point Estimation

- Bias, Variance and MSE
- CAN estimator
- MLE

Confidence Interval

Hypothesis Testing

- Duality of Confidence Interval and Hypothesis Tests

p-value

Point Estimation

Point Estimator

- A **point estimator** of the parameter θ is a function from the sample space to the parameter space:

$$\hat{\theta}(\cdot) : \mathcal{X} \to \Theta.$$

- **Estimator** vs. **Estimate**: The former is a function, while the latter is the realized value of this function based on an observed sample:
  - before observing any data, the sample data is *random*, therefore $\hat{\theta}(\mathbf{X})$ is random
  - once we observe the data $\mathbf{X} = \mathbf{x}$, the estimate $\hat{\theta}(\mathbf{x})$ is a number
- Examples of point estimators: the OLS estimator, the sample mean, etc.

Sample Mean

Let $X_1, \ldots, X_n$ be a random sample drawn independently from a population. $\mathbb{E}[X_i] = \mu$, $\mathbb{V}[X_i] = \sigma^2$ are the population mean and variance. Assume σ^2 is known, μ is unknown and is the parameter of interest.

We consider the sample mean estimator $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Is it a good estimator? What about ...

- median? $\bar{X} - 1$? 5?
- sample mean, but with truncation such as

$$\frac{1}{n}\sum_{i=1}^{n}[X_i \mathbb{1}(|X_i| \leq 10) + 10\text{sgn}(X_i)\mathbb{1}(|X_i| > 10)]?$$

Bias and Variance

- The **bias** of an estimator $\hat{\theta}$ of θ is defined to be

$$\mathbb{E}_\theta[\hat{\theta}(\mathbf{X})] - \theta.$$

  An **unbiased estimator** is one for which the bias is zero.

  Is the sample mean $\bar{X}$ unbiased?

- The variance of an estimator is $\mathbb{V}_\theta[\hat{\theta}(\mathbf{X})]$.

  What's the variance of $\bar{X}$? What about the few alternatives we considered?

Mean Squared Error

The **mean squared error** (MSE) of an estimator $\hat{\theta}$ for the point estimation of θ is

$$\mathbb{E}_\theta[(\hat{\theta} - \theta)^2].$$

Note:

- MSE is commonly used, but it is just one of many ways to evaluate estimators.
- Exercise: $MSE(\hat{\theta}) = \mathbb{V}_\theta(\hat{\theta}) + Bias(\hat{\theta})^2$. MSE captures both the precision of the estimator (variance over different samples) as well as the accuracy (biasedness).
- An estimator that is biased (many Bayesian estimators) but more precise might be preferable to one that is unbiased but fluctuates wildly.

Mean Squared Error - Exercise

1. Derive the MSE for the sample mean estimator $\bar{X}$
2. Consider an alternative estimator for μ, the linear shrinkage estimator $\hat{\mu} = \omega \bar{X}$, where $\omega \in [0, 1]$. Find the Bias, Variance and MSE of $\hat{\mu}$, and compare its MSE vs the MSE for $\bar{X}$.

Asymptotic Behaviors of Sample Mean

If we increase the sample size, i.e., n goes to ∞, will the sample mean estimator get better? In what sense?

Let $\bar{X}_n$ denote the sample mean estimator for a sample of size n. Notice that $\mathbb{V}[\bar{X}_n] = \sigma^2/n$ goes to 0 as $n \to \infty$.

Is $\bar{X}_n$ getting closer to μ? For some $\epsilon > 0$, consider

$$\mathbb{P}(|\bar{X}_n - \mu| > \epsilon).$$

Example: Compute this probability for $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ in closed form, and then its limit when $n \to \infty$.

Probability Inequalities

Theorem

Markov's Inequality: *Let X be an integrable non-negative random variable. Then for any $t > 0$,*

$$\mathbb{P}(X > t) \leq \frac{E[X]}{t}.$$

Theorem

Chebyshev's Inequality: *For any square integrable random variable X with $\mathbb{E}[X] = \mu$ and $\mathbb{V}[X] = \sigma^2$,*

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

Convergence in Probability

Let $\{X_1, \ldots, X_n\}$ be a sequence of random variables, X be a random variable. We say X_n **converges** to X **in probability**, if for all $\epsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

Usually denote as $X_n \xrightarrow{p} X$.

The Law of Large Numbers (LLN)

Let $\{X_1, \ldots, X_n\}$ be a sequence of independently and identically distributed (i.i.d.) square integrable random variables with $E[X_i] = \mu$, let $\bar{X}_n$ be the sample mean.

The Weak Law of Large Numbers (WLLN) states:

$$\bar{X}_n \xrightarrow{p} \mu \text{ as } n \to \infty.$$

That is, for any $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|\bar{X}_n - \mu| > \epsilon) = 0.$$

Exercise: use Chebyshev's inequality to prove WLLN.

There are many versions of LLN, see this link as a start if interested.

LLN Application - Monte Carlo Methods

An application of LLN is the use of Monte Carlo methods as numerical approximation for expectation of functions.

Consider a random variable $X \sim f$, for some distribution f. We are interested in $\mathbb{E}[h(X)]$ for some function $h(\cdot)$.

If we are able to sample from f, it's natural to consider the following estimator:

1. sample $x_1, \ldots, x_n \overset{i.i.d.}{\sim} f$
2. $\hat{h}_n(X) = \frac{1}{n}\sum_{i=1}^n h(x_i)$.

By WLLN, if $\mathbb{E}[h(X)]$ exists, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|\hat{h}_n(X) - \mathbb{E}[h(X)]| > \epsilon) = 0,$$

i.e., $\hat{h}_n(X)$ is a consistent estimator.

Monte Carlo Methods - Probability

This generalizes into approximation for probability and integrals.

For any event A, let $\mathbb{1}_A(X)$ denote the indicator function where

$$\mathbb{1}_A(X) = \begin{cases} 1 & \text{if } X \in A \\ 0 & \text{otherwise.} \end{cases}$$

Notice that

$$\mathbb{P}(X \in A) = \mathbb{E}[\mathbb{1}_A(X)].$$

Therefore we can use Monte Carlo methods to approximate probabilities without a closed form.

Monte Carlo Methods - Probability Example

Let $X \sim N(1, 3)$, use Monte Carlo simulations to estimate $\mathbb{P}[X \leq 2]$.

Notice that $\mathbb{P}[X \leq 2] = \mathbb{E}[\mathbb{1}(X \leq 2)]$. Therefore we can approximate it as follows:

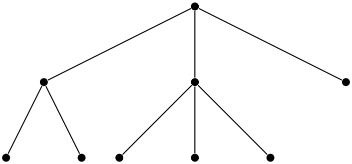
1. sample $x_1, \ldots, x_n \overset{i.i.d.}{\sim} N(1, 3)$
2. compute

$$\hat{\mathbb{P}}[X \leq 2] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(x_i \leq 2).$$

See R demo.

Monte Carlo Methods - Advanced Example (Optional)

A Poisson(λ) **branching process** is a tree where each node has a random number of children drawn from a Poisson(λ) distribution.



Commonly used in ecology, genetics, particle physics, random graphs, etc. Want to know

$$P[\text{Tree survives forever}]$$

as a function of λ. Can approximate by simulating many trees. See *R* demo.

Monte Carlo Methods - Integration

Consider the integral

$$\int_0^2 x^2 dx.$$

There is a closed form solution from calculus, which is

$$\int_0^2 x^2 dx = \frac{1}{3}x^3\Big|_0^2 = \frac{8}{3}.$$

Alternatively, use Monte Carlo simulation. Notice

$$\int_0^2 x^2 dx = 2\int_0^2 \frac{1}{2}x^2 dx = 2\mathbb{E}[X^2], \quad X \sim Unif(0,2).$$

Monte Carlo Methods - Integration Example

Therefore we can approximate $\int_0^2 x^2 dx$ as follows:

1. sample $x_1, \ldots, x_n \overset{i.i.d.}{\sim} Unif(0, 2)$
2. compute

$$\frac{1}{n} \sum_{i=1}^{n} x_i^2.$$

See R demo.

Monte Carlo Methods - Example

Think about how to use Monte Carlo simulations to approximate the following integral:

$$\int_{-\infty}^{\infty} \frac{x}{1+x^2} dx.$$

Is this a good estimator?

Hint: the PDF for a standard Cauchy distribution is

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

Recap

We considered the sample mean as a point estimator for the population mean, and evaluated its performance.

- For finite sample:
  - Bias, Variance and MSE
- Asymptotically, we have shown that for an integrable random variable X,

$$\bar{X}_n \xrightarrow{p} \mathbb{E}[X],$$

i.e., the sample mean is a *consistent* estimator.

Consistency

A sequence of estimator $\{\hat{\theta}_n\}$ is **consistent** for estimating θ if

$$\hat{\theta}_n \xrightarrow{p} \theta \text{ as } n \to \infty.$$

- The probability that a consistent estimator is more than ϵ away from the parameter is vanishingly small, for any $\epsilon > 0$.
- Consistent estimators get close to the parameter with high probability as sample size increases.

Anything else we can say about the sample mean?

The Central Limit Theorem (CLT)

Let $\{X_1, \ldots X_n\}$ be a sequence of i.i.d. random variables with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$.

Let $\bar{X}_n$ denote the sample mean, then as $n \to \infty$, the sequence of random variables $\sqrt{n}(\bar{X}_n - \mu)$ converges in distribution to $N(0, \sigma^2)$:

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

Equivalently:

$$\lim_{n \to \infty} \mathbb{P}\left(\sqrt{n}\frac{(\bar{X}_n - \mu)}{\sigma} \leq c\right) = \Phi(c).$$

where $\Phi(\cdot)$ denotes the CDF for a standard normal.

Convergence in Distribution

A sequence $\{X_1, X_2, \dots\}$ of random variables is said to **converge in distribution** to a random variable X if

$$\lim_{n\to\infty} F_n(t) = F(t),$$

for every $t \in \mathbb{R}$ at which F is continuous, where F_n and F denote the CDFs for X_n and X.

Note that CLT applies regardless of the underlying distribution of the data as long as variance is finite and samples are i.i.d.

See R demo.

CAN Estimators

We have shown that for i.i.d. data with finite mean μ and variance σ^2, the sequence of sample mean estimator $\{\bar{X}_n\}$ is **C**onsistent for the population mean and **A**symptotically **N**ormal, i.e.,

$$\bar{X}_n \xrightarrow{p} \mu, \quad \sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

Hence it's a *CAN estimator*.

More formally, Let $\{\hat{\theta}_n\}$ be a consistent sequence of estimators for θ and

$$\sqrt{n}\frac{(\hat{\theta}_n - \theta)}{\sigma} \xrightarrow{d} N(0, 1).$$

Then we say $\hat{\theta}_n$ is a **CAN estimator** for θ, with asymptotic variance σ^2.

Likelihood and Maximum Likelihood Estimator (MLE)

Consider the statistical model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Let $f(x \mid \theta)$ be the PDF of P_θ. For $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta$, the joint density is

$$p(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

The **likelihood function** is the joint density evaluated at realized value $\mathbf{x}$ as a function of parameter θ:

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

The **log-likelihood function** is

$$\ell(\theta; \mathbf{x}) = \log L(\theta; \mathbf{x}) = \sum_{i=1}^{n} \log f(x_i \mid \theta).$$

MLE - Example

A **maximum likelihood estimator** (MLE) of the parameter θ based on a random sample $\mathbf{X}$ is

$$\hat{\theta}(\mathbf{X}) = \text{argmax}_{\theta \in \Theta} L(\theta; \mathbf{X}).$$

It's the parameter value at which the likelihood function attains maximum, i.e., the *most probable* value.

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ with σ^2 known. Find the MLE of μ.

MLE - Exercise

Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ with both μ and σ^2 unknown. Find the MLEs for both μ and σ^2. Are they unbiased?

Properties of MLE

We have seen that for finite samples, MLE is not necessarily unbiased. What about other properties?

- Is it efficient (small MSE) for finite sample, and asymptotically?
- Is it consistent?
- Is it asymptotically normal?

Properties of MLE

Under regularity conditions, the MLE is

1. consistent,
2. asymptotically normal,
3. efficient, or asymptotically optimal, and
4. equivariant to transformation of parameters.

We will only briefly go through the properties here. These materials will be properly covered later in the Statistical Inference class.

Properties of MLE

Let $\hat{\theta}_n$ be the MLE for parameter θ based on a sample of size n. Under regularity conditions,

1. $\hat{\theta}_n$ is consistent for θ, i.e., $\hat{\theta}_n \xrightarrow{p} \theta$.
2. $\hat{\theta}_n$ is asymptotically normal,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\left(0, \frac{1}{I(\theta)}\right),$$

where $I(\theta)$ is the **Fisher Information**,

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ell(\theta; X)}{\partial \theta^2}\right] = \mathbb{V}\left[\frac{\partial \ell(\theta; X)}{\partial \theta}\right].$$

Properties of MLE

3. $\hat{\theta}_n$ is asymptotically optimal in the following sense. By the **Cramer-Rao lower bound**, for any other unbiased estimator $\tilde{\theta}$ of θ, and any $\theta \in \Theta$,

$$\mathbb{V}[\tilde{\theta}] > \frac{1}{nI(\theta)},$$

   i.e., MLE achieves the smallest asymptotic variance among all unbiased estimators.

4. Let $\tau = g(\theta)$ be a function of θ, the MLE for τ is

$$\hat{\tau}_n = g(\hat{\theta}_n).$$

MLE - Optional Exercise

Let's use an exercise to gain a better understanding of the equivariance property of MLE.

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} Bin(\theta)$. Find the MLE of θ.

Now rewrite the model in terms of

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} Bin\left(\frac{e^{\psi}}{1+e^{\psi}}\right).$$

Find the MLE of ψ and relate it to the MLE of θ.

Confidence Interval

Interval Estimator

Recall point estimation gives a single value estimate of the parameter of interest based on sample data. In contrast, interval estimation gives a range of plausible values.

- An **interval estimator** C_n of a parameter $\theta \in \Theta \subset \mathbb{R}$ is a set-valued function: $\mathcal{X} \to 2^{\Theta}$, mapping from the sample space to subsets of the parameter space.
- Specifically $C_n = (L(\mathbf{X}), U(\mathbf{X}))$, where $L(\cdot)$ and $U(\cdot)$ are functions: $\mathcal{X} \to \mathbb{R}$, $L(\mathbf{X}) \leq U(\mathbf{X})$.

Level $(1-\alpha)$ Confidence Interval

We call C_n a level $(1-\alpha)$ **confidence interval**(CI), or we say C_n has (1-α) coverage if

$$\mathbb{P}(\mathbf{X} \in C_n(\mathbf{X}) \mid \theta) \geq 1-\alpha, \quad \text{for all } \theta \in \Theta.$$

Note: this is not a probability statement about θ. The parameter θ is fixed. It's the data vector $\mathbf{X}$ and the confidence interval $C_n(\mathbf{X})$ that are random.

- A common but **WRONG** interpretation: there's $(1-\alpha)$ probability that the parameter θ is in the interval C_n.
- **CORRECT** interpretation: if we repeat the experiment many times, and construct intervals $C_n(\mathbf{X}_i)$ for each sample $\mathbf{X}_i$. $(1-\alpha)$ % of these intervals are expected to contain θ.

Confidence Interval - Exercise

Suppose that $\mathbf{X}$ is a random sample from a distribution with parameter θ, and $[L(\mathbf{X}), U(\mathbf{X})]$ is a 95% CI of θ. If we observe $\mathbf{X} = \mathbf{x}$, which of the following statements is correct?

A The probability that $\theta \in [L(\mathbf{x}), U(\mathbf{x})]$ is 0.95;

B The probability that $\theta \in [L(\mathbf{x}), U(\mathbf{x})]$ is either 1 or 0.

Confidence Interval - Example

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ with σ^2 known. We want to construct a level $(1-\alpha)$ CI for μ.

Many possible ways! First consider using the Markov's inequality: for integrable non-negative random variable X, and any $t > 0$,

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}[X]}{t}.$$

As $|\bar{X}_n - \mu|$ is non-negative, we have

$$\mathbb{P}(|\bar{X}_n - \mu| > c) \leq \frac{\mathbb{E}[|\bar{X}_n - \mu|]}{c} \leq \frac{\mathbb{E}[(\bar{X}_n - \mu)^2]^{1/2}}{c} = \frac{\sigma}{\sqrt{n}c}.$$

For the second inequality, recall we showed $\|X\|_p \leq \|X\|_q$ for all $0 < p < q < \infty$.

Confidence Interval - Example

In order to construct a level $(1-\alpha)$ CI, we want

$$\mathbb{P}(|\bar{X}_n - \mu| > c) \leq \frac{\sigma}{\sqrt{n}c} \leq \alpha \implies c \geq \frac{\sigma}{\sqrt{n}}\frac{1}{\alpha}.$$

Set $c = \sigma/(\sqrt{n}\alpha)$ (we want narrow CIs!), a level (1-α) CI for μ is

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

Notice that $\sigma/\sqrt{n}$ is the standard deviation of $\bar{X}_n$. Taking $\alpha = 0.05$, this CI is 40 times the standard deviation!!
Can we do better?

Confidence Interval - Exercise

Still let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ with σ^2 known.

Following a similar approach as in the example,

1. Construct a level $(1-\alpha)$ CI for μ using the Chebyshev inequality: If $\mathbb{E}[X] = \mu$ and $\mathbb{V}[X] = \sigma^2$,
$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$
2. Construct an exact level $(1-\alpha)$ CI for μ based on the distribution of $\bar{X}_n$.
3. Compare the widths of these CIs using $\alpha = 0.05$.

Confidence Interval - Example Cont.

Same setup, but with σ^2 unknown. Let s^2 denote the sample variance estimator. Then

$$T_{n-1} = (\bar{X}_n - \mu)/(s/\sqrt{n}) \sim t_{n-1}$$

is a pivotal quantity, i.e., its distribution is independent of μ. Therefore a $(1-\alpha)$ confidence interval of μ is given by

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

where $t_{df,p}$ is the $p \times 100\%$th quantile of a student-t distribution with df degrees of freedom.

Confidence Interval - Optional Example

Consider the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim N_2(\mathbf{0}, \sigma^2\mathbf{I})$, with $\mathbf{Y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n\times 2}$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^T \in \mathbb{R}^2$, σ^2 unknown.

Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)^T$ be the OLS estimator, $\mathbf{P}$ be the orthogonal projection matrix onto the column space of $\mathbf{X}$,

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix}.$$

Construct an exact level (1-α) CI for β_1.

Hint: recall we have shown $\hat{\beta} \sim N_2(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$. Check the *Distributions* slides, what would be an appropriate estimator for σ^2 for us to apply the approach in the last example?

Asymptotic Level $(1-\alpha)$ Confidence Interval

What about general situations when we do not have a pivotal quantity? One approach: make use of asymptotic normality of MLE.

- Let $\hat{\theta}_n$ be the MLE of parameter θ. We know

$$\sqrt{nI(\theta)}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0,1).$$

- As $\hat{\theta}_n$ is consistent, by the *Continuous Mapping Theorem*,

$$\sqrt{nI(\hat{\theta}_n)}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0,1).$$

- Therefore an asymptotic level $(1-\alpha)$ confidence interval for θ is

$$\left(\hat{\theta}_n - Z_{1-\frac{\alpha}{2}}[nI(\hat{\theta}_n)]^{-\frac{1}{2}},\ \hat{\theta}_n + Z_{1-\frac{\alpha}{2}}[nI(\hat{\theta}_n)]^{-\frac{1}{2}}\right).$$

Hypothesis Testing

Hypothesis Testing

We use **Hypothesis Testing** to decide whether some hypothesis formulated is likely to be correct.

Consider the statistical model $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$. Here's the most common setup.

- Let $\{\Theta_H, \Theta_K\}$ be a partition of the parameter space such that $\Theta_H \cap \Theta_K = \emptyset$, $\Theta_H \cup \Theta_K = \Theta$.
- We want to test the null hypothesis

$$H : \theta \in \Theta_H,$$

against the alternative hypothesis

$$K : \theta \in \Theta_K.$$

Test Errors and Power Function

- **Type I Error** and **Type II Error**:

		Decision	
		Accept H	Reject H
Truth	H	Correct decision	Type I Error
	K	Type II Error	Correct decision

- Let R denote the **rejection region** for a test.
  - Probability of Type I Error: $\mathbb{P}(\mathbf{X} \in R|H)$.
  - Probability of Type II Error: $\mathbb{P}(\mathbf{X} \in R^c|K) = 1 - \mathbb{P}(\mathbf{X} \in R|K)$.
- A **level-α test** is one such that $\mathbb{P}(\mathbf{X} \in R|H) \leq \alpha$.

Hypothesis Testing Procedure

Testing procedure based on sample data $\mathbf{x}$:

1. Identify a test statistics $T(\mathbf{x})$ which distinguishes H and K (typically larger T indicates H is less likely to be true).
2. Find the null distribution of $T(\mathbf{X})$ and the critical value c for a proper level-α test
3. Accept H if $T(\mathbf{x}) < c$, reject otherwise.

Hypothesis Testing - Example

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ with σ^2 known. We want to test:

$$H : \mu = 0 \quad \text{vs} \quad K : \mu > 0.$$

- A natural test statistics is $T(\mathbf{x}) = \bar{x}$, the sample mean. The larger the sample mean, the less likely H is true.
- Distribution of $\bar{X}$ under H (i.e., the null distribution) is

$$\bar{X} \sim N(0, \sigma^2/n).$$

- Therefore to set up a level α test, we can reject when

$$\bar{X} > Z_{1-\alpha} \frac{\sigma}{\sqrt{n}}.$$

Duality of Confidence Interval and Hypothesis Tests

There is one-to-one correspondence between level (1-α) confidence intervals and level α hypothesis tests.

Confidence interval to test

- Let $C : \Omega \to 2^{\Theta}$ be a level (1-α) confidence interval for parameter $\theta \in \Theta$.
- A level α test procedure for $H : \theta = \theta_0$ is to accept H if $\theta_0 \in C(\mathbf{X})$.

Test to confidence interval

- Let $A(\theta_0)$ be the acceptance region of a level α test for $H : \theta = \theta_0$.
- A level (1-α) confidence interval is $C(\mathbf{x}) = \{\theta \in \Theta : \mathbf{x} \in A(\theta)\}$, i.e., all parameter values that won't be rejected after observing $\mathbf{x}$.

p-value

p-value

A **p-value** $p(\mathbf{X})$ is the probability of obtaining test results at least as extreme as the observed statistics, assuming the null hypothesis is correct. It measures evidence against the null hypothesis.

For $H : \theta = \theta_0$, it is typically set up as:

$$p(\mathbf{x}) = \mathbb{P}(T(\mathbf{X}) \geq T(\mathbf{x}) | H),$$

where:

- $T(\mathbf{X})$: A test statistic (such as $\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$), i.e., a function of the random data $\mathbf{X}$. Typically larger values indicate deviation from H.
- $T(\mathbf{x})$: $T(\cdot)$ evaluated at the observed data $\mathbf{x}$.

p-value - Example

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug, and recording their response time.

The neurologist knows that the response time for a rat not injected with the drug follows a normal distribution with a mean response time of 1.2 seconds.

The mean of the 100 injected rats' response times is 1.05 seconds with a sample standard deviation of 0.5 seconds.

Do you suggest that the neurologist conclude that the drug has an effect on response time?

Reference Guide

- *Statistical Inference* - Casella and Berger
- *A First Course in Bayesian Statistical Methods* - Hoff
- *All of Statistics* - Wasserman

Acknowledgement

Past contributors:

- Jordan Bryan, PhD student
- Brian Cozzi, MSS alumni
- Michael Valancius, MSS alumni
- Graham Tierney, PhD student
- Becky Tang, PhD student

This set of slides made reference to

- 2021S STA732 course materials
- 2020S STA532 course materials