Introduction to Statistics Note (Mid-term Exam)

2024 Spring Semester

21 CST H3Art

Chapter 1: Looking at Data——Distributions

1.1 Data

Statistics（统计学） is the science of learning from data.

Cases（案例） are the objects described by a set of data. Cases may be customers, companies, experimental subjects, or other objects.

A **variable（变量）** is a special characteristic of a case.

A **label（标签）** is used in some data sets to provide additional information about a variable.

Different cases can have different **values** of a variable.

A **categorical（分类的）** variable places each case into one of several groups, or categories. Sometimes also referred to as a nominal variable (i.e., used for naming). Example: first language, ethnicity.

A **quantitative（定量的）** variable takes numerical values for which arithmetic operations such as adding and averaging make sense.

The **distribution** of a variable tells us the values that a variable takes and how often it takes each value.

1.2 Displaying Distributions with Graphs

The distribution of a **categorical variable（分类变量）** lists the categories and gives the **count** or **percent** of individuals who fall into each category:

- **Pie charts（饼图）** show the distribution of a categorical variable as a “pie” whose slices are sized by the **counts** or **percents** for the categories.
- **Bar graphs（条形图：条一般分开来）** represent categories as bars whose heights show the category counts or percents.

The distribution of a **quantitative variable（定量变量）** tells us what values the variable takes on and how often it takes those values:

- **Histograms（直方图：条一般连在一起）** show the distribution of a quantitative variable by using bars. The height of a bar represents the number of individuals whose values fall within the corresponding class.
- **Stemplots（茎叶图）** (or **stem-and-leaf plots**) separate each observation into a stem and a leaf that are then plotted to display the distribution while maintaining the original values of the variable.

A distribution is **symmetric** if the right and left sides (or tails) of the graph are approximately mirror images of each other.

A distribution is **skewed to the right** (**right-skewed**) if the right side of the graph (containing the half of the observations with larger values) is much longer than the left side.

It is skewed to the **left** (**left-skewed**) if the left side of the graph is much longer than the right side.

An important kind of deviation is an **outlier（离群值）**. Outliers are observations that lie outside the overall pattern of a distribution.

1.3 Describing Distributions with Numbers

The most common measure of center (or central tendency) is the **arithmetic average（算术平均值）**, or **mean（均值）**.

$$\bar{x} = \frac{\text{sum of observations}}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum x_i$$

The **median** is the midpoint of a distribution, the number such that half of the observations are smaller and the other half are larger.

The **mean** and **median** of a roughly **symmetric** distribution are close together. If the distribution is exactly **symmetric**, the mean and median are exactly **the same**.

In a **skewed distribution**, the **mean** is usually **farther out in the long tail（均值趋向于长尾方向）** than is the median.

A measure of center alone can be **misleading**. And a useful numerical description of a distribution requires **both a measure of center and a measure of spread** (or variability), so we introduce **quartiles（分位点）**:

- Arrange the observations in increasing order and locate the **median** M.
- The **first quartile** Q_1 is the median of the observations located to the left of the median in the ordered list.
- The **third quartile** Q_3 is the median of the observations located to the right of the median in the ordered list.
- The **interquartile range** (IQR) is defined as: $\mathrm{IQR} = Q_3 – Q_1$.

The **five-number summary** of a distribution consists of the **smallest observation**, the **first quartile**, the **median**, the **third quartile**, and the **largest observation**, written in order from smallest to largest:

$$\min \ Q_i \ M \ Q_3 \ \max$$

The median and quartiles divide the distribution roughly into quarters. This leads to a new way to display quantitative data, the **boxplot**:

- Draw and label a number line that includes the range of the distribution.
- Draw a central box from Q_1 to Q_3.
- Note the **median** M inside the box.
- Extend lines (whiskers) from the box out to the **minimum** and **maximum** values that are **not outliers**.

The $1.5 \times \mathrm{IQR}$ Rule for Outliers: Call an observation an outlier if it falls more than $1.5 \times \mathrm{IQR}$ **above the third quartile** or **below the first quartile**.

The most common measure of spread looks at how far each observation is from the mean. This measure is called the **standard deviation**(s_x).

$$\text{variance} = s_x^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum (x_i - \bar{x})^2$$

$$\text{standard deviation} = s_x = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$$

s_x measures spread about the mean and should be used only when the **mean** is **the measure of center（仅当均值为测量值的中心时适用）**.

s_x is **not** resistant to outliers.**（无法对抗离群值）**

Numerical summaries do not fully describe the shape of a distribution.

Changing the Unit of Measurement:

- **Multiplying** each observation by a positive number b multiplies both **measures of center (mean, median)** and **spread** (IQR, s_x) by b.
- **Adding** the same number a (positive or negative) to each observation adds a to **measures of center** and to **quartiles**, but it does **not** change measures of spread (IQR, s_x).

1.4 Density Curves and Normal Distributions

A **density curve（密度曲线）** is a curve that:

- is always **on** or **above** the horizontal axis
- has an **area** of exactly 1 underneath it

Distinguishing the **Median** and **Mean** of a **Density Curve**:

- The **median** of a density curve is the **equal-areas point**——the point that divides the area under the curve in half.
- The **mean** of a density curve is the **balance point**, that is, the point at which the curve would balance if made of solid material.
- The **mean** of a **skewed curve** is **pulled away from the median** in the direction of the **long tail**. **（均值在倾斜曲线中总是更偏向长尾一端）**

A **Normal distribution** is described by a **Normal density curve**. Any particular Normal distribution is completely specified by two numbers: its **mean** μ and **standard deviation** σ.

- The **mean** of a Normal distribution is the center of the symmetric Normal curve.
- We abbreviate the Normal distribution with mean μ and standard deviation σ as $N(\mu, \sigma)$.

The 68-95-99.7 Rule:

- Approximately 68% of the observations fall within σ of μ.
- Approximately 95% of the observations fall within 2σ of μ.
- Approximately 99.7% of the observations fall within 3σ of μ.

If a variable x has a distribution with **mean** μ and **standard deviation** σ, then the standardized value of x, or its z-score, is:

$$z = \frac{x - \mu}{\sigma}$$

The **standard Normal distribution（标准正态分布）** is the Normal distribution with **mean** 0 and **standard deviation** 1.

One way to assess if a distribution is indeed approximately Normal is to plot the data on a **Normal quantile plot**.

- If the distribution is indeed Normal, the plot will show a **straight line**, indicating a **good match** between the data and a Normal distribution.
- Systematic **deviations** from a straight line indicate a **non-Normal distribution**.
- **Outliers** appear as points that are far away from the overall pattern of the plot.

Chapter 2: Looking at Data——Relationships

2.1 Relationships

Two variables measured on the same cases are **associated（相关）** if knowing the value of one of the variables tells you something that you would not otherwise know about the value of the other variable.

A **response variable（响应变量/因变量）** measures an outcome of a study.

An **explanatory variable（解释变量/自变量）** explains or causes changes in the response variable.

Certain characteristics of a data set are key to exploring the relationship between two variables. These should include the following:

- Cases
- Label
- Categorical or quantitative
- Values
- Explanatory or response

2.2 Scatterplots

The most useful graph for displaying the relationship between **two quantitative variables** is a **scatterplot（散点图）**.

A scatterplot shows the relationship between **two quantitative variables** measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis.

How to Make a Scatterplot

- Decide which variable should go on each axis. If a distinction exists, plot the **explanatory variable** on the x axis and the **response variable** on the y axis.
- Label and scale your axes.
- Plot individual data values.

Two variables are **positively associated** when above-average values of one tend to accompany above-average values of the other, and when below-average values also tend to occur together.

Two variables are **negatively associated** when above-average values of one tend to accompany below-average values of the other, and vice-versa.

2.3 Correlation

The **correlation** r measures the strength of the linear relationship between two quantitative variables. Using the notation explained on pp. 103–104 in the text:

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Properties of Correlation:

- r is always a number **between** -1 **and** 1.
- $r > 0$ indicates a **positive** association.
- $r < 0$ indicates a **negative** association.
- Values of r near 0 indicate a very **weak linear relationship**.
- The strength of the linear relationship increases as r moves away from 0 toward -1 or 1.
- The extreme values $r = -1$ and $r = 1$ occur only in the case of a perfect linear relationship.
- **Correlation** makes no distinction between explanatory and response variables.**（相关性不区分因变量和自变量）**
- r has no units and does not change when we change the units of measurement of x, y, or both.
- Positive r indicates positive association between the variables, and negative r indicates negative association.
- Correlation requires that **both variables be quantitative**.
- Correlation does **not describe curved relationships** between variables, no matter how strong the relationship is.
- The correlation r is **not resistant**; it can be strongly affected by a few outlying observations.

- Correlation is **not a complete summary** of two-variable data.

2.4 Least-Squares Regression

A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes. We can use a regression line to **predict** the value of y for a given value of x.

Regression equation:

$$\hat{y} = b_0 + b_1 x$$

- x is the value of the **explanatory variable**.
- $\hat{y}$ is the **predicted value** of the response variable for a given value of x.
- b_1 is the **slope（斜率）**, the amount by which y changes for each oneunit increase in x.
- b_0 is the **intercept（截距）**, the value of y when $x = 0$.

Least-Squares Regression Line （最小二乘回归线）(LSRL):
The least-squares regression line of y on x is the line that minimizes the sum of the squares of the vertical distances of the data points from the line.

The **square of the correlation**, r^2, is the fraction of the variation in values of y that is explained by the least-squares regression of y on x.

- r^2 is called the **coefficient of determination（决定系数）**.

2.5 Cautions About Correlation and Regression

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line:

$$\begin{aligned} \text{residual} &= \text{observed } y - \text{predicted } y \\ &= y - \hat{y} \end{aligned}$$

Outliers in the x direction are often **influential** for the least-squares regression line, meaning that the removal of such points would **markedly change the equation of the line**.

Cautions About Correlation and Regression:

- Both describe **linear** relationships.
- Both are affected by **outliers**.
- Always plot the data **before interpreting**.
- Beware of **extrapolation（外推值）**.
  - Use caution in predicting y when x is **outside the range** of observed x's.
- Beware of **lurking variables（潜在变量）**.
  - These have an **important effect** on the relationship among the variables in a study, but are not included in the study.
- **Correlation** does **not** imply **causation（因果关系）**!

2.7 The Question of Causation

Association, however **strong**, does **NOT** imply **causation**.

- **Common Response（共同反应）**: The observed relationship between the variables can be explained by a **lurking variable（潜在变量）**.
- **Confounding（混淆）**: Two variables are confounded when their effects on a response variable **cannot be distinguished** from each other.

We can evaluate the association using the following **criteria**:

- The association is **strong（强大的）**.
- The association is **consistent（始终一致的）**.
- Higher **doses（剂量）** are associated with stronger responses.
- **Alleged cause（声称的原因）** precedes the effect.
- The alleged cause is **plausible（合理的）**.

Chapter 3: Producing Data

3.1 Sources of Data

Anecdotal data（轶事数据） represent individual cases that often come to our attention because they are striking in some way. "The plural of **anecdote** is not evidence."

Sample surveys are a special type of **designed experiment** that usually aim to discover the opinions of people on certain topics. In a sample survey, a **sample** of individuals is selected from a larger **population** of individuals.

The **distinction** between **observational study（观察研究）** and **experiment（实验）** is one of the most important in statistics:

- An **observational study** observes individuals and measures variables of interest but **does not attempt to influence the responses**.
- An **experiment** deliberately **imposes some treatment** on individuals to measure their responses.

Observational studies of the effect of one variable on another often **fail** because of **confounding** between the explanatory variable and one or more **lurking variables**:

- **Confounding（混淆）** occurs when two variables are associated in such a way that **their effects** on a response variable **cannot be distinguished** from each other.
- A **lurking variable（潜在变量）** is a variable that is **not among the explanatory or response variables** in a study but that may influence the response variable.

3.2 Design of Experiments

An **experiment** is a study in which we actually do something (a **treatment**) to people, animals, or objects (the **experimental units**) to observe the **response**.

An **experimental unit** is the smallest **entity** to which a **treatment is applied**.

- When the units are **human beings**, they are often called **subjects（主体）**.

The **explanatory variables** in an experiment are often called **factors（因子）**.

A specific condition applied to the individuals in an experiment is called a **treatment**.

Many laboratory experiments operate as follows:

$$\text{Experimental Units} \rightarrow \text{Treatment} \rightarrow \text{Measure Response}$$

Outside the laboratory, **badly designed experiments** often **yield worthless results（产生无价值的结果）** because of **confounding（混淆）**.

In a **comparative experiment**, **comparison alone** isn't enough. If the treatments are given to groups that differ greatly, **bias（偏差）** will result. The **solution** to the problem of bias is **random assignment（随机分配）**.

In a **completely randomized** design, the treatments are assigned to all the experimental units completely by chance. Some experiments may include a **control group（控制变量组）** that **receives an inactive treatment（接受不明显的对待）** or an existing **baseline treatment**.

How to randomly choose n individuals from a group of N:

- We first label each of the N individuals with a number (typically from 1 to N, or 0 to $N-1$).
- Imagine writing the whole numbers from 1 to N on separate pieces of paper. Now put all the numbers in a hat.
- Mix up the numbers and randomly select one.
- Mix up the remaining $N-1$ numbers and randomly select one of them.
- Continue in this way until we have our sample of n numbers.

Principles of Experimental Design:

- **Control** for lurking variables that might affect the response, most simply by comparing two or more treatments. **（控制潜在变量）**
- **Randomize**: Use chance to assign experimental units to treatments. **（使对待随机化）**
- **Replication**: Use enough experimental units in each group to reduce chance variation in the results. **（复制实验，足够多的实验才有说服力）**

An observed effect so large that it would rarely occur by chance is called **statistically significant（统计显著性）**.

A **statistically significant association（统计上的显著关联）** in data from a **well-designed experiment** does imply **causation（因果关系）**.

In a **double-blind experiment（双盲实验）**, neither the **subjects** nor those **who interact with them** and measure the response variable **know which treatment a subject received**. **（被试者和实验者均不知道他们做了什么操作）**

A **matched pairs（配对）** design is a **randomized blocked experiment** in which each block consists of a **matching pair of similar experimental units**.

A **block** is **a group of experimental units** that are known before the experiment to be **similar** in some way that is expected to affect the response to the treatments.

3.3 Sampling Design

The **population** in a statistical study is the **entire group of individuals** about which we want information.

A **sample** is the part of the population from which we actually collect information.

The design of a **sample** is **biased** if it **systematically favors** certain outcomes. **（系统性地偏向某些结果）**

A **voluntary response sample（自愿响应样本）** consists of people who choose themselves by responding to a general appeal. Voluntary response samples often show bias because people with strong opinions (often in the same direction) may be more likely to respond.

A **simple random sample （简单随机抽样）(SRS)** of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance to be the sample actually selected.

A **probability sample** is a sample chosen by chance. We must know what samples are possible and what chance, or probability, each possible sample has.

To select a **stratified random sample（分层随机抽样）**, first classify the population into groups of similar individuals, called **strata（阶层）**. Then choose a separate SRS in each stratum and combine these SRSs to form the full sample.

Systematic sample（系统抽样） arranges the units in the population in a certain order, **determine the selection interval** according to the sample size requirements, and select a unit at certain intervals.

Good sampling technique includes the art of reducing all sources of **error**:

- **Undercoverage（覆盖不足）** occurs when some groups in the population are left out of the process of choosing the sample.
- **Nonresponse（无响应）** occurs when an individual chosen for the sample can't be contacted or refuses to participate.
- A systematic pattern of incorrect responses in a sample survey leads to **response bias（响应偏差）**.
- The **wording of questions（问题措辞）** is the most important influence on the answers given to a sample survey.

3.4 Toward Statistical Inference

A **parameter（参数）** is a number that describes some characteristic of the population.**（描述总体）**

A **statistic（统计量）** is a number that describes some characteristic of a sample.**（描述样本）**

We write μ **(the Greek letter mu)** for the **population mean** and σ for the population standard deviation.

We write $\bar{x}$ **(x-bar)** for the **sample mean** and s for the sample standard deviation.

The **population distribution** of a variable is the distribution of values of the variable among all individuals in the population.

The **sampling distribution** of a **statistic** is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

Bias（偏差） concerns the **center of the sampling distribution**. A **statistic** used to estimate a **parameter** is **unbiased**.

- To reduce bias, use **random sampling（随机抽样）**.

The **variability（变化性）** of a **statistic** is described by **the spread of its sampling distribution**.

- To reduce variability of a statistic from an SRS, use a **larger sample（更大的样本）**.

The process of drawing **conclusions about a population** on the basis of **sample data** is called **inference**.

Chapter 4: Probability: The Study of Randomness

4.1 Randomness

We call a phenomenon **random** if individual outcomes are **uncertain** but there is nonetheless a regular distribution of outcomes in a **large number of repetitions**.

The **probability** of any outcome of a chance process is the proportion of times the outcome would occur in a very long series of repetitions.

4.3 Random Variables

A numerical variable that describes the outcomes of a chance process is called a **random variable**.

The **probability distribution** of a random variable gives its possible values and their probabilities.

We **standardize** normal data by calculating **z-scores** so that any Normal curve $N(\mu, \sigma)$ can be transformed into the standard Normal curve $N(0, 1)$.

$$z = \frac{(x - \mu)}{\sigma}$$

4.4 Means and Variances of Random Variables

Draw independent observations at random from any population with finite mean μ. The **law of large numbers（大数定律）** says that, as the number of observations drawn increases, the sample mean of the observed values gets closer and closer to the mean μ of the population.**（观测的样本均值越来越接近总体均值μ）**

Chapter 5: Sampling Distributions

5.1 The Sampling Distribution of a Sample Mean

Mean of a sampling distribution of a sample mean:
There is no tendency for a sample mean to fall systematically above or below μ, even if the distribution of the raw data is skewed. Thus, the sample mean is an **unbiased estimate（无偏估计量）** of the population mean μ.

$$\hat{\mu} = \mu$$

Standard deviation of a sampling distribution of a sample mean:
The standard deviation of the sampling distribution measures how much the sample statistic varies from sample to sample. It is smaller than the standard deviation of the population by a factor of $\sqrt{n}$.

$$s_x = \frac{\sigma}{\sqrt{n}}$$

The Central Limit Theorem（中心极限定理）:
Draw an SRS of size n from any population with mean μ and standard deviation σ. The **central limit theorem (CLT)** says that when n is sufficiently large, the sampling distribution of the **sample mean is approximately Normal**, specifically,

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

5.2 Sampling Distributions for Proportions

Choose an SRS of size n from a population with p as the true proportion of success $\rightarrow$ it follows that the population standard deviation is $\sigma = \sqrt{p(1-p)}$

- The mean of the sampling distribution of $\hat{p}$ is $\mu_{\hat{p}} = p$
- The standard deviation of the sampling distribution of $\hat{p}$ is $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$

As n **increases**, the sampling distribution **becomes approximately Normal**.

For sufficiently large n:

$$\hat{p} \sim N(p, \sqrt{p(1-p)/n})$$

Chapter 6: Introduction to Inference

6.1 Estimating with Confidence

Statistical inference（统计推断） provides methods for drawing conclusions about a population from sample data.

Confidence Interval（置信区间）:

A **level C confidence interval** for a parameter has two parts:

- An **interval** calculated from the data, which has the form:

$$\text{estimate} \pm \text{margin of error}$$

- A confidence level C, where C is the probability that the interval will capture the true parameter value in repeated samples. In other words, the confidence level is the success rate for the method.

Choose an SRS of size n from a population having unknown mean μ and known standard deviation σ. A level C **confidence interval** for μ is:

$$\bar{x} \pm z \times \frac{\sigma}{\sqrt{n}}$$

The critical value z is found from the **standard Normal distribution**.

The confidence level Cdetermines the value of z, the **margin of error** also depends on z.

$$m = z \times \frac{\sigma}{\sqrt{n}}$$

The **margin of error（误差范围）** gets smaller when:

- z gets **smaller** (the same as a lower confidence level C).
- σ is **smaller**. It is easier to pin down μ when σ is smaller.
- n gets **larger**. Since n is under the square root sign, we must take four times as many observations to cut the margin of error in half.

The confidence interval for a population mean will have a specified margin of error m when the sample size is:

$$m = z \times \frac{\sigma}{\sqrt{n}} \leftrightarrow n = \left(\frac{z \times \sigma}{m} \right)^2$$

Some Cautions:

- The data should be an **SRS（简单随机抽样）** from the population.
- The **confidence interval** and **sample size formulas** are **not** correct for **other sampling methods**.
- Inference **cannot** rescue **badly produced data**.
- Confidence intervals are not resistant to outliers.
- If n is **small (<15)** and the population is not Normal, the true confidence level will be **different** from C.
- The standard deviation σ of the population **must be known**.
- The margin of error in a confidence interval **covers only random sampling errors**!

6.2 Tests of Significance

A **test of significance（显著性检验）** is a formal procedure for comparing observed data with a claim (also called a **hypothesis（假设）**) whose truth we want to assess.

We express the results of a significance test in terms of a probability, called the **P-value**, that measures how well the data and the claim agree.

A **significance test** starts with a careful statement of the claims we want to compare.

- The claim tested by a statistical test is called the **null hypothesis（零假设）(H_0)**. The test is designed to assess the strength of the evidence against the null hypothesis.
- The claim about the population for which we're trying to find evidence is the **alternative hypothesis（备选假设）(H_a)**.
  - The alternative is **one-sided** if it states either that a parameter is **larger** than the null hypothesis value, or **smaller** than the null hypothesis value.
  - It is **two-sided** if it states that the parameter is **different** from the null value.

The **null hypothesis** H_0 states the claim that we seek to disprove. The probability that measures the strength of the evidence **against** a null hypothesis is called a **P-value**.

- **Small P-values** are evidence **against** H_0 because they say that the observed result is unlikely to occur when H_0 is true.
  - P-value small $\rightarrow$ reject H_0 $\rightarrow$ conclude H_a (in context)
- **Large P-values fail** to give convincing evidence against H_0 because they say that the observed result is likely to occur by chance when H_0 is true.
  - P-value large $\rightarrow$ fail to reject H_0 $\rightarrow$ cannot conclude H_a (in context)

We can compare the **P-value** with a fixed value that we regard as decisive, called the **significance level**, we write it as α:

- P-value < α $\rightarrow$ reject H_0 $\rightarrow$ conclude H_a (in context)
- P-value ≥ α $\rightarrow$ fail to reject H_0 $\rightarrow$ cannot conclude H_a (in context)

z test for a population mean:
Draw an SRS of size n from a Normal population that has unknown mean μ and known standard deviation σ. To test the null hypothesis that μ has a specified value,

$$H_0 : \mu = \mu_0$$

calculate the one-sample z statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

In terms of a variable Z having the standard Normal distribution, the P-value for a test of H_0 against

$$H_a : \mu > \mu_0 \text{ is } P(Z \geq z)$$

$$H_a : \mu < \mu_0 \text{ is } P(Z \leq z)$$

$$H_a : \mu \neq \mu_0 \text{ is } 2P(Z \geq |z|)$$

6.4 Power and Inference as a Decision

If we **reject** H_0 when H_0 **is true**, we have committed a **Type I error**.

If we **fail to reject** H_0 when H_0 **is false**, we have committed a **Type II error**.

If you insist on a **smaller significance level** (such as 1% rather than 5%), you have to take a **larger sample**.

A significance test makes a **Type II error** when it fails to reject a null hypothesis that really is false. There are many values of the parameter satisfying the alternative hypothesis. We can calculate the probability that a test does reject H_0 when any specific alternative is true. This probability is called the **power** of the test against that specific alternative.（当显着性检验未能拒绝确实错误的原假设时，就会出现 II 类错误。 满足备择假设的参数值有很多。 我们可以计算当任何特定替代方案为真时测试拒绝H_0的概率。 该概率称为针对该特定替代方案的测试**功效**）

If you insist on **higher power** (such as 99% rather than 90%), you will need a **larger sample**.

The Common Practice of Testing Hypotheses:

- State H_0 and H_a as in a test of significance.
- Think of the problem as a decision problem, so the probabilities of **Type I** and **Type II** errors are relevant.
- Consider only tests in which the probability of a **Type I error** is **no greater** than a specified α.
- Among these tests, select a test that makes the probability of a **Type II error** as **small** as possible.

Chapter 7: Inference for Distributions

7.1 Inference for the Mean of a Population

When the sampling distribution of $\bar{x}$ is close to Normal, we can find probabilities involving $\bar{x}$ by standardizing:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

When we don't know σ, we can estimate it using the sample standard deviation s_x, our statistic has a new distribution called a **t distribution**.

$$t = \frac{\bar{x} - \mu}{s_x/\sqrt{n}}$$

There is a different t distribution for each sample size, specified by its **degrees of freedom （自由度）(df)**, the one-sample t statistic has the **t distribution** with degrees of freedom df = $n–1$.

The One-Sample t Interval for a Population Mean:

Choose an SRS of size n from a population having unknown mean . A level C **confidence interval** for μ is:

$$\bar{x} \pm t \times \frac{s_x}{\sqrt{n}}$$

where t is the **critical value** for the $t(n–1)$ distribution.

The **margin of error** is:

$$t \times \frac{s_x}{\sqrt{n}}$$

The One-sample t Test:

Choose an SRS of size n from a large population that contains an unknown mean μ. To test the hypothesis $H_0 : \mu = \mu_0$, compute the one-sample t statistic:

Matched Pairs t Procedures:

To compare the responses to the two treatments in a matchedpairs design, find the difference between the responses within each pair. Then apply the one-sample t procedures to these differences.

7.2 Comparing Two Means

When data come from two random samples or two groups in a randomized experiment, the statistic $\bar{x_1} - \bar{x_2}$ is our best guess for the value of $\mu_1 - \mu_2$.

When the two samples are independent of each other, the **standard deviation** of the statistic $\bar{x_1} - \bar{x_2}$ is:

$$s_{\bar{x_1}-\bar{x_2}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

We standardize the observed difference to obtain a t statistic:

$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

When the Random, Normal, and Independent conditions are met, a level C confidence interval for $(\mu_1 - \mu_2)$ is:

$$(\bar{x_1} - \bar{x_2}) \pm t \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where t is the critical value at confidence level C for the t distribution with degrees of freedom either gotten from technology or equal to the smaller of $n_1 - 1$ and $n_2 - 1$.

Approximate Distribution of the Two-Sample t Statistic:

The distribution of the two-sample t statistic is very close to the t distribution with degrees of freedom given by:

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

This approximation is accurate when both sample sizes are 5 or **larger**.

Pooled Two-Sample Procedures:

degrees of freedom: $n_1 + n_2 - 2$

Suppose both populations are Normal and they have the same standard deviations. The pooled estimator of σ^2 is:

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

A level C confidence interval for $\mu_1 - \mu_2$ is:

$$(\bar{x_1} - \bar{x_2}) \pm t \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where the degrees of freedom for t are $n_1 + n_2 - 2$

To test the hypothesis $H_0 : \mu_1 = \mu_2$ against a **one-sided** or a **two-sided** alternative, compute the pooled two-sample t statistic for the $t(n_1 + n_2 - 2)$ distribution.

$$t = \frac{\bar{x_1} - \bar{x_2}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$