# Foundations of Statistics

Yuzu Sensei

2024-11-23

# Table of contents

# Preface

All about Statistics.

# 1 Basics

Knuth (1984)

## 1.1 Descriptive Statistics

population deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

standard deviation of a sample

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n-1}}$$

is descriptive statistics, which is a description of the variation in measurements. However, the standard error of the mean is descriptive of the random sampling process, which is a probabilistic statement about how the sample size will provide a better bound on estimates of the population mean, in light of the central limit theorem.

For a sample of size $n$,

standard deviation of the sample mean

$$s.d(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

but since $\sigma$ is unknow, we use standard error of the sample mean

$$s.e(\bar{X}) = \frac{s}{\sqrt{n}}$$

Put simply, the **standard error** of the sample mean is an estimate of how far the sample mean is likely to be from the population mean, whereas the **standard deviation** of the

sample is the degree to which individuals within the sample differ from the sample mean.[9] If the population standard deviation is finite, the standard error of the mean of the sample will tend to zero ($s.e \to 0$ )with increasing sample size, because the estimate of the population mean will improve, while the standard deviation of the sample $s$ will tend to approximate the population standard deviation $\sigma$ as the sample size increases.

## 1.2 Confidence interval

Suppose $X_1, ..., X_n$ is an independent sample from a population $Normal(\mu, \sigma^2)$.

Sample mean

$$\bar{X} = \frac{X_1 + ... + X_n}{n}$$

Sample variance

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

Then

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{\bar{X} - \mu}{s.e}$$

s.e is the standard error of the sample mean

## 1.3 Randomisation

How can we use sampled data to inform us about the population from which the data are drawn?

Suppose that $X_1, X_2, ..., X_n$ represent a random sample from a distribution with mean $\mu$ , then

$$E[X_1 + X_2 + ... + X_n] = E[X_1] + E[X_2] + ... + E[X_n] = n\mu.$$

Since $X_1, X_2, ..., X_n$ are independent (hence uncorrelated), we have

$$\text{Var}(X_1 + X_2 + ... + X_n) = \text{Var}(X_1) + \text{Var}(X_2) + ... + \text{Var}(X_n) = n\sigma^2$$

$$\begin{aligned}\text{sd}(X_1 + X_2 + ... + X_n) &= \sqrt{\text{Var}(X_1 + X_2 + ... + X_n)} \\ &= \sqrt{\text{Var}(X_1) + \text{Var}(X_2) + ... + \text{Var}(X_n)} \\ &= \sqrt{n\sigma^2} = \sqrt{n}\sigma\end{aligned}$$

We now have the mean and standard deviation of the sample mean $\bar{X}$

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

$$\begin{aligned}\text{sd}(\bar{X}) = \text{sd}\left(\frac{\sum X_i}{n}\right) &= \sqrt{\text{Var}\left(\frac{\sum X_i}{n}\right)} \\ &= \sqrt{\frac{1}{n^2}\text{Var}(\sum X_i)} \\ &= \sqrt{\frac{1}{n^2}n\sigma^2} \\ &= \frac{\sigma}{\sqrt{n}}\end{aligned}$$

where $\mu$ is the population mean and $\sigma$ is the population standard deviation.

population standard deviation

$$\sigma = \sqrt{\text{Var}(X)} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n}}$$

sample standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N - 1}}$$

However, in practice nobody knows $\sigma$. So we use standard error

$$\text{se}(\bar{x}) = \frac{s_x}{\sqrt{n}},$$

where $s_x$ is the sample standard deviation.

Note that here we use $\text{se}(\bar{x})$ to indicate it is a fixed value for a particular sample. The corresponding random variable will be denoted by $\text{se}(\bar{X})$.

The sampling distribution of the sample mean is $\bar{X} \sim \text{Normal}(\mu_{\bar{X}} = \mu, \sigma_{\bar{X}} = \sigma/\sqrt{n})$. Thus, in the long run, the observed sample mean falls within $\pm 2\text{sd}(\bar{X})$ of the population mean $\mu$ for approximately 95% of samples taken.

As the sample size $n$ increases (the number of random samples), the estimate $\bar{X}$ becomes more precise because $\text{sd}(\bar{X})$ becomes smaller.

$$Z = \frac{\bar{X} - E[\bar{X}]}{\text{sd}(\bar{X})} = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1)$$

## 1.4 Central limit theorem (CLT)

$$\sqrt{n}(\bar{X} - u) \sim \text{Normal}(0, \sigma^2)$$

> **Note**
>
> If you repeatedly sample a random variable a large number of times, the distribution of the sample mean will approach a normal distribution regardless of the initial distribution of the random variable.
> The sample mean $\bar{X}$ is approximately Normally distributed in large samples.
> Heaviness of the tails of the distribution and lack of symmetry are important factors in slowing down the CLT effect.
> Diagnostic: one should always plot the sample data. If the data look as though they may have come from a distribution that isn't extremely non-normal, we can feel much more confident about calculations based on a Normal approximation with moderate-sized samples.

Why this is useful? Because in practice we never know the exact form of the distribution we are sampling from (e.g., can be exponential, trangular, uniform…), but CLT tells us we can apply normal-distribution theory for means from large samples even when the original distribution is not Normal.

sample size for CLT to work: it depends. If a distribution is close in shape to the Normal, the CLT works fast (e.g, 4-12). For an exponential distribution, it requires 30-50 (depends on the degree of skewness).

## 1.5 Bias

bias= $E(\hat{\Theta}) - \theta$

## 1.6 Precision

The precision of the estimate is a measure of how variable the estimator is in repeated sampling.

A precise estimate is one that is subject to very little sampling variation.

measure of precision: standard deviation of the sampling distribution of the estimator sd($\hat{\Theta}$)

However, the actual standard deviation of the sampling distribution is unknown. So we use an estimate of the actual standard deviation - standard error se($\hat{\theta}$) as the measure of the precision of our estimate.

Smaller standard errors correspond to more precise estimates.

> **Precision**
>
> The standard error of any estimate $\hat{\theta}$
> se($\hat{\theta}$) estimates the variability of $\hat{\theta}$ values in repeated sampling
> it is a measure of precision of $\hat{\theta}$ .

Note:

- An estimate can be biased but precise.
- An estimate can be precise but biased

### 1.6.1 Accuracy

An accurate estimator is one that generally gives estimates that are close to the parameter estimated so that it will have low bias and high precision.

### 1.6.2 Sample proportion

The sample proportion $\hat{p}$ estimates the population proportion $p$.

The number of people reported having seen illegal drugs $Y \sim \text{Binomial}(n, p)$. We have

$$E[Y] = np, \text{Var}(Y) = np(1-p)$$

The sample proportion random variable $\hat{P} = Y/n$.

$$E[\hat{P}] = E[Y/n] = \frac{1}{n}E[Y] = p, \text{sd}(\hat{P}) = \sqrt{\text{Var}(\hat{P})} = \sqrt{\text{Var}(Y/n)} = \sqrt{\frac{1}{n^2}np(1-p)} = \sqrt{\frac{p(1-p)}{n}}$$

By CLT, we have $\hat{P} \sim \text{Normal}(E[\hat{P}], \text{sd}(\hat{P})^2)$.

However, in reality, we wouldn't know the population proportion $p$, so we want to use a sample proportion $\hat{p}$ to estimate this unknown $p$. Thus, we use standard error

$$\text{se}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

### 1.6.3 Standard error for a difference between independent estimates

$$\text{se}(\hat{\theta_1} - \hat{\theta_2}) = \sqrt{\text{se}(\hat{\theta_1})^2 + \text{se}(\hat{\theta_2})^2}$$

$$\text{se}(\hat{p_1} - \hat{p_2}) = \sqrt{\text{se}(\hat{p_1})^2 + \text{se}(\hat{p_2})^2} = \text{se}(\hat{p}) = \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

$$H_0 : \hat{p} = p_0$$
$$H_1 : \hat{p} \neq p_0$$

$$Z = \frac{\hat{p} - p_0}{\text{sd}(\hat{p})} = \frac{\hat{p} - p_0}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1)$$

# 2 Hypothesis Testing

X: the number of heads when we toss a coin 10 times

$H_0$: $X \sim$ Binomial (n=10,p=0.5) (If the coin is fair)

We observe $X = 9$, and get p-value 0.021

## 2.1 P-value

p-value is the probability of observing something at least as extreme as our observation, if the null hypothesis is true.

For example, in the above example, we calculate the p-value of 0.021 as a measure of the strength of evidence against the hypothesis that $p = 0.5$ (fair coin).

# 3 Hypothesis Testing, p-value and Significance

## 3.1 P-value Interpretation

p-value =0.021

We have some evidence that $p \neq 0.5$ (the coin is not fair)

it only tells us that in our sample, we have evidence that $p$ is different from 0.5

but it does not mean the difference between true $p$ and 0.5 is large, or the difference between true $p$ and 0.5 is important in practical terms.

So what we get is " Substantial evidence of a difference", not "Evidence of a substantial difference".

Most people agree that p-value is a useful measure of the strength of evidence against the null hypothesis.

The smaller the p-value, the stronger the evidence against the null hypothesis.

As a rule of thumb, we consider the p-values of 0.05 and less start to suggest that the null hypothesis is doubtful.

## 3.2 Statistical Significance

The result of a hypothesis test is significant at the 5% level if the p-value is less than 0.05.

The chance of seeing what we did see (9 heads out of 10 tosses), or more, is less than 5% if the null hypothesis is true.

Note that 5% of the time, we will get a p-value <0.05 when the null hypothesis is TRUE!

A small p-value does Not mean that the null hypothesis is definitely wrong.

## 3.3 Hypothesis testing Example

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

## 3.4 Z-test vs T-test

when sample size is large and population variance $\sigma^2$ is known,

$$Z = \frac{\bar{X} - E[\bar{X}]}{\text{sd}(\bar{X})} = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1)$$

When sample size is small and population variance $\sigma^2$ is unknown, we use the sample variance $s^2$ instead

The t-test is parametrized by the degrees of freedom, which refers to the number of independent observations in a dataset, denoted by $\nu = n - 1$

$$T = \frac{\bar{X} - E[\bar{X}]}{\text{se}(\bar{X})} = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim \text{Student}(df = n - 1),$$

where $s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$.

The additional variability of T is reflected in its distribution being flatter and having longer tails than the standard Normal.

T-distribution is similar to the normal distribution in appearance but has larger tails. This means that extreme events happen with greater frequency than the modeled distribution would predict.

# 4 Upcoming Content

I will think about it.

# References

Knuth, Donald E. 1984. “Literate Programming.” *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.