Chapter 7

Sampling distribution

7.1 Samples and statistics

We model real-world uncertain events with random variables. We have also introduced various distributions suitable to model different kinds of events. However, we never observe the full distribution or the true parameters of the assumed distribution. Instead, we only observe a sample of that random variable. We can only infer the properties of the distribution from a limited sample. For example, suppose we model the hight of an Asian women with a normal distribution. But we never know exactly what the mean and variance are. We can only observe a sample of the distribution.

In statistics, the conceptual distribution F is called the **population distribution**, or just the **population**.[1] It is tempting to think of the population as all the observations (e.g. all the population on the planet), but this is not exactly correct. The population distribution is more of a mathematical abstraction or an assumption. Suppose we are modeling the height of human being, even if we have all the observations on the planet, that does not include the people that have died or yet to be born. Thus, it is still a sample of the assumed distribution.

A collection of random variables $\{X_1, X_2, \ldots, X_n\}$ is a **random sample** from the population F if X_i are **independent and identically distributed (*i.i.d*)**

[1] This section is based on Bruce Hansen's *Probability and Statistics for Economists*.

with distribution F. What we mean by *i.i.d* is that $X_1, \ldots, X_n$ are mutually independent and have exactly the same distribution $X_i \sim F$. Survey sampling is an useful metaphor to understand random sampling, in which we randomly select a subset of the population with equal probability. The **sample size** n is the number of individuals in the sample.

A **data set** is a collection of numbers, typically organized by observation. We sometimes call a data set also as a sample. But it should not be confused with the random sample defined above. As the former is a collection of random variables, whereas the latter is one **realization** of the random variables.

Typically, we will use X without the subscript to denote a random variable or vector with distribution F, X_i with a subscript to denote a random observation in the sample, and x_i or x to denote a specific or realized value.

The problem of **statistical inference** is to learn about the underlying process — the population distribution or data generating process — by examining the observations. In most cases, we assume the population distribution and want to learn about the its parameters (e.g. μ and σ^2 in the normal distribution). As a convention, we use greek letters to denote population parameters.

A **statistic** is a function of the random sample $\{X_1, X_2, \ldots, X_n\}$. Recall that there is a distinction between random variables and their realizations. Similarly there is a distinction between a statistic as a function of a random sample — and is therefore a random variable as well — and a statistic as a function of the realized sample, which is a realized value. When we treat a statistic as random we are viewing it is a function of a sample of random variables. When we treat it as a realized value we are viewing it as a function of a set of realized values. One way of viewing the distinction is to think of "before viewing the data" and "after viewing the data". When we think about a statistic "before viewing" we do not know what value it will take. From our vantage point it is unknown and random. After viewing the data and specifically after computing and viewing the statistic the latter is a specific number and is therefore a realization. It is what it is and it is not changing. The randomness is the process by which the data was generated — and the understanding that if this process were repeated the sample would be different and the specific realization would be therefore different. The distribution of a statistic is called the **sampling distribution**, since it is the distribution induced by sampling.

An **estimator** $\hat{\theta}$ for a population parameter θ is a statistic intended to infer θ.

It is conventional to use the hat notation $\hat{\theta}$ to denote an estimator. Note that $\hat{\theta}$ is a statistic and hence also a random variable. We call $\hat{\theta}$ an **estimate** when it is a specific value (or realized value) calculated in a specific sample.

A standard way to construct an estimator is by the analog principle. The idea is to express the parameter θ as a function of the population F, and then express the estimator $\hat{\theta}$ as the analog function in the sample.

For example, suppose we want to construct an estimator for the population mean $\mu = E(X)$. By definition, if each value of X is of equal probability, μ is simply the average. By analogy, we construct the **sample mean** as $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. It is conventional to denote a sample average by the notation "X bar". Because it is an estimator for μ, we also denote it as $\hat{\mu} = \bar{X}_n$. Note that from different samples we calculate different estimates. In one sample, $\hat{\mu} = 6.5$; in another sample, $\hat{\mu} = 6.7$. All of them are erroneous estimate of the true parameter μ. The question is therefore how close they are to the true parameter. To answer this question, we need to study the distribution of the sample mean.

7.2 Law of large numbers

We now introduce two important theorems describing the behavior of the sample mean as the sample size grows. Throughout this section and the next, we assume $X_1, X_2, \ldots, X_n$ are i.i.d RVs drawn from a population with mean μ and variance σ^2. The sample mean is defined as

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

As we have discussed previously, the sample mean is itself a random variable with mean and variance:

$$\begin{aligned} E(\bar{X}_n) &= \frac{1}{n}E(X_1 + \cdots + X_n) = \frac{1}{n}(E(X_1) + \cdots + E(X_n)) = \mu, \\ Var(\bar{X}_n) &= \frac{1}{n^2}Var(X_1 + \cdots + X_n) \overset{iid}{=} \frac{1}{n^2}(Var(X_1) + \cdots + Var(X_n)) = \frac{\sigma^2}{n}. \end{aligned}$$

The law of large numbers (LLN) says that as n grows, the sample mean $\bar{X}_n$ converges to the true mean μ.

Theorem 7.1 (Strong law of large numbers)**.** *The sample mean $\bar{X}_n$ converges to the true mean μ point-wise as $n \to \infty$, with probability* 1*. In other words, the event $\bar{X}_n \to \mu$ has probability* 1*.*

Theorem 7.2 (Weak law of large numbers)**.** *For all $\epsilon > 0$, $P(|\bar{X}_n - \mu| > \epsilon) \to 0$ as $n \to \infty$. (this is known as converge in probability).*

We don't need a rigorous proof here. But an intuitive proof is obvious. As $n \to \infty$, $Var(\bar{X}_n) = \frac{\sigma^2}{n} \to 0$. The random variable $\bar{X}_n$ becomes fixed at μ as n becomes large. Thus, it converges to μ in a probabilistic sense.

It seems that the LLN just states the obvious. But it has wide applications in daily time that you might not even realize. What it says is essentially this: the uncertainty at the individual level becomes certain when aggregating together; the risks that are unmanageable at the individual level becomes manageable collectively. Think about a rare disease, it happens at 1 out of a million probability. For each individual, no one knows if they will get the disease or not. But as the sample size gets large, suppose we have one billion population, the LLN says the sample mean will be very close the true mean. That is, there will be almost surely 1000 people being infected by the disease. We provide two more examples.

Example 7.1 (Lottery)**.** A lottery company is designing a game with a 6-digit format. Each time someone buys a ticket, they receive a randomly generated 6-digit number. Only one number will win the grand prize of 10 million dollars. What should the company charge per ticket to break even?

Solution: The probability of winning the game is $p = 1/10^6$. Suppose the company has sold n tickets. The price for each ticket is x. The revenue for the company is therefore xn. By the LLN, the cost of the company should be very close to $10^7 np$. The break even point is $xn = 10^7 np$. So $x = 10^7 p = 10$. Therefore, if the company sells each ticket above 10 dollars. The business is surely profitable as long as n is large. If the company is a monopoly, it can reap as much profit as it desires as long as they know the basic probability theory! The same can be said about gambling companies.

Example 7.2 (Insurance)**.** Insurance is anther great application of the LLN. It is essentially the same as the the lottery game but most people do not realize it. Suppose there is a disease with infection rate of 1 out of 1 million. The medical

expenditure to cure the disease is 10 million dollars. How much the insurance company should charge per customer to cover this disease?

Solution: The solution is essentially the same as above. Suppose the premium for the insurance product is x. The revenue of the company by selling the premium is xn. The cost is — when one customer is infected, the company has to pay the medical cost —$10^7 np$. The break even price for the insurance premium is thus 10 dollars.

What is the implication of this insurance? Without the insurance, each individual either chooses to set aside 10 million dollars pre-cautiously for the disease (if he is rich enough) or be exposed to the risk completely uncovered. The insurance product enables everyone to get covered at a cost of just 10 dollars. It is a typical example that the unmanageable risk at the individual level becomes manageable collectively.

7.3 Central limit theorem

The LLN shows the convergence of the sample mean to the population mean. What about the entire sample distribution? This is addressed by the central limit theorem (CLT), which, as its name suggests, is a limit theorem of **central importance** in statistics.

The CLT states that for large n, the distribution of $\bar{X}_n$ after standardization approaches a standard Normal distribution, regardless of the underlying distribution of X_i. By **standardization**, we mean that we subtract μ, the expected value of $\bar{X}_n$, and divide by $\sigma/\sqrt{n}$, the standard deviation of $\bar{X}_n$.

Theorem 7.3 (Central limit theorem)**.** *As $n \to \infty$,*

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \to N(0,1) \text{ in distribution.}$$

In other words, the CDF of the left-hand side approaches the CDF of the standard normal distribution.

Proof. We will prove the CLT assuming the MGF of the X_i exists, though the theorem holds under much weaker conditions. Without loss of generality let

$\mu = 1, \sigma^2 = 1$ (since we standardize it anyway). We show that the MGF of $\sqrt{n}\bar{X}_n = (X_1 + \cdots + X_n)/\sqrt{n}$ converges to the MGF of the $N(0,1)$.

The MGF of $N(0,1)$ is

$$\begin{aligned}
E(e^{tX}) &= \int_{-\infty}^{\infty} e^{tx} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2 + tx} dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2 + \frac{1}{2}t^2} dx \\
&= e^{\frac{t^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2} dx \\
&= e^{t^2/2}
\end{aligned}$$

Compute the MGF of $\sqrt{n}\bar{X}_n$:

$$\begin{aligned}
E(e^{\sqrt{n}\bar{X}_n}) &= E(e^{t(X_1+\cdots+X_n)/\sqrt{n}}) \\
&= E(e^{tX_1/\sqrt{n}}) E(e^{tX_2/\sqrt{n}}) \cdots E(e^{tX_n/\sqrt{n}}) \\
&= \left[E(e^{tX_i/\sqrt{n}}) \right]^n \qquad \text{since } i.i.d \\
&= \left[E\left(1 + \frac{tX_i}{\sqrt{n}} + \frac{t^2 X_i^2}{2n} + o(n^{-1}) \right) \right]^n \\
&= \left[1 + \frac{t}{\sqrt{n}} E(X_i) + \frac{t^2}{2n} E(X_i^2) + o(n^{-1}) \right]^n \\
&= \left[1 + \frac{t^2}{2n} + o(n^{-1}) \right]^n \\
&= \left[1 + \frac{t^2/2}{n} + o(n^{-1}) \right]^n \\
&\to e^{t^2/2} \qquad \text{as } n \to \infty
\end{aligned}$$

Therefore, the MGF of $\sqrt{n}\bar{X}_n$ approaches the MGF of the standard normal. Since MGF determines the distribution, the distribution of $\sqrt{n}\bar{X}_n$ also approaches the standard normal distribution. □

The CLT tells us about the limiting distribution of $\bar{X}_n$ as $n \to \infty$. That means,

we can reasonably approximate the distribution $\bar{X}_n$ with normal distribution when n is a finite large number —

$$\bar{X}_n \approx N(\mu, \sigma^2/n) \quad \text{for large } n.$$

The Central Limit Theorem was first proved by Pierre-Simon Laplace in 1810. Let's take a moment to admire the generality of this result. The distribution of the individual X_i can be anything in the world, as long as the mean and variance are finite. This does mean the distribution of X_i is irrelevant, however. If the distribution is fairly close to normal, the result would hold for smaller n. If the distribution is far away from normal, it would take larger n to converge.

The CLT gives the distribution of the sample mean regardless of the underlying distribution. This allows to assess the "quality" of the sample mean — how close it is to the true mean. The LLN tells us the larger the sample, the closer the sample mean to the population mean. The CLT tells us the distribution of the sample mean for sample size n. For smaller n, the distribution is more spread-out (a normal distribution with large σ^2); hence the uncertainty is huge, other values are more likely. For larger n, the uncertainty is reduced, most values would be centered around the true mean. We will delve deeper into this when we get to hypothesis testing.

Example 7.3. Suppose that a fair coin is tossed 900 times. Approximate the probability of obtaining more than 395 heads.

Solution: Let $H = \sum_{i=1}^{900} X_i$ be the number of heads, where $X_i \sim \text{Bern}(\frac{1}{2})$. We could compute the probability by

$$P(H > 495) = \sum_{k=496}^{900} \binom{900}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{900-k}$$

But this is quite tedious. Because $n = 900$ is reasonably large, we can apply the CLT:

$$\frac{1}{n}\sum_{i=1}^{900} X_i \sim N(\mu, \sigma^2/n) \quad \text{or}$$
$$\sum_{i=1}^{900} X_i \sim N(n\mu, n\sigma^2)$$

We know $\mu = E(X_i) = \frac{1}{2}$, $\sigma^2 = Var(X_i) = \frac{1}{4}$. Thus $H \sim N(450, 225)$. Therefore,

$$P(H > 495) = 1 - P(H \leq 495) \approx 1 - \Phi\left(\frac{495 - 450}{15}\right) = 0.0013.$$

7.4 Estimator accuracy

This section introduces some measures regarding the accuracy of an estimator.

Definition 7.1. The **bias** of an estimator $\hat{\theta}$ of a parameter θ is

$$\text{Bias}[\hat{\theta}] = E(\hat{\theta}) - \theta.$$

We say that an estimator is **biased** if its sampling is incorrectly centered. We say that an estimator is **unbiased** is the bias is zero.

Theorem 7.4. *$\bar{X}_n$ is **unbiased** for $\mu = E(x)$ if $E(X) < \infty$.*

Proof.

$$E(\bar{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu.$$

□

Theorem 7.5. *If $\hat{\theta}$ is an unbiased estimator of θ, then $\hat{\beta} = a\hat{\theta} + b$ is an unbiased estimator of $\beta = a\theta + b$.*

But obtaining an unbiased estimator is not always as straightforward as it seems. Consider the sample variance as an estimator for the population variance. By the analog principle, the sample variance should be

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\right)^2 \\ &= \tilde{\sigma}^2 - (\bar{X}_n - \mu)^2\end{aligned}$$

We know that

$$E(\tilde{\sigma}^2) = \frac{1}{n}\sum_{i=1}^{n} E(X_i - \mu)^2 = \sigma^2$$

Thus, if we compute the bias of this estimator:

$$\begin{aligned}E[\hat{\sigma}^2] &= \sigma^2 - \frac{\sigma^2}{n} = \left(1 - \frac{1}{n}\right)\sigma^2 \\ \text{Bias}[\hat{\sigma}^2] &= -\frac{\sigma^2}{n} \neq 0\end{aligned}$$

Therefore, the estimator $\hat{\sigma}^2$ is a biased estimator for σ^2! To correct the bias, we divide the sample sum of squares by $(n-1)$.

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

It is straightforward to see that s^2 is an unbiased estimator for σ^2. We call s^2 the **bias-corrected variance estimator**.

Theorem 7.6. *s^2 is an unbiased estimator for σ^2 if $E(X^2) < \infty$.*

Definition 7.2. The mean square error of an estimator $\hat{\theta}$ for θ is

$$\text{MSE}[\hat{\theta}] = E\left[(\hat{\theta} - \theta)^2\right].$$

By expanding the square we find that

$$\begin{aligned}
\text{MSE}[\hat{\theta}] &= E\left[(\hat{\theta}-\theta)^2\right] \\
&= E\left[(\hat{\theta}-E[\hat{\theta}]+E[\hat{\theta}]-\theta)^2\right] \\
&= E\left[(\hat{\theta}-E[\hat{\theta}])^2\right] + 2E(\hat{\theta}-E[\hat{\theta}])(E[\hat{\theta}]-\theta) + (E[\hat{\theta}]-\theta)^2 \\
&= Var[\hat{\theta}] + (\text{Bias}[\hat{\theta}])^2.
\end{aligned}$$

Thus the MSE is the variance plus the squared bias. The MSE as a measure of accuracy combines the variance and bias.

Theorem 7.7. *For any estimator with a finite variance, we have*

$$MSE[\hat{\theta}] = Var[\hat{\theta}] + (Bias[\hat{\theta}])^2.$$

Definition 7.3. An estimator is **consistent** if $\text{MSE}[\hat{\theta}] \to 0$ as $n \to \infty$.

Bias is the property of an estimator for finite samples. Consistency is the property of an estimator when the sample size gets large. It means that for any given data distribution, there is a sample size n sufficiently large such that the estimator $\hat{\theta}$ will be arbitrarily close to the true value θ with high probability. In practice, we usually do not know how large this n has to be. But it is a desirable property for an estimator to be considered a "good" estimator.

For unbiased estimator, MSE is solely determined by the variance of the estimator. Recall that the variance for the sample mean is $Var(\bar{X}_n) = \sigma^2/n$. But this is not a very useful formula because the it depends on unknown parameter σ^2. We need to replace these unknown parameters by estimators. To put the latter in the same units as the parameter estimate we typically take the square root before reporting. We thus arrive at the following concept.

Definition 7.4. A **standard error** of an estimator $\hat{\theta}$ is defined as

$$SE(\hat{\theta}) = \hat{V}^{1/2}$$

where $\hat{V}$ is the estimator for $Var[\hat{\theta}]$.

Definition 7.5. The **standard error** for $\bar{X}_n$ is

$$SE(\bar{X}_n) = \frac{s}{\sqrt{n}}$$

where s is the bias-corrected estimator for σ.

Note the difference between **standard error** and **standard deviation**. Standard deviation describes the dispersion of a distribution. Standard error is the standard deviation of an *estimator*. It indicates the "precision" of the estimator, thereby carrying a sense of "error". The smaller the standard error, the more precise the estimator.

7.5 Confidence intervals

Confidence intervals provide a method of adding more information to an estimator $\hat{\theta}$ when we wish to estimate an unknown parameter θ. We can find an interval (A, B) that we think has high probability of containing θ. The length of such an interval gives us an idea of how closely we can estimate θ.

Definition 7.6. A $100(1-\alpha)\%$ **confidence interval (CI)** for θ is an interval $[L(\theta), U(\theta)]$ such that the probability that the interval contains the true θ is $(1-\alpha)$.

Due to randomness we rarely seek a confidence interval with 100% coverage as this would typically need to be the entire parameter space. Instead we seek an interval which includes the true value with reasonably high probability. Standard choices are $\alpha = 0.05$ and 0.10, corresponding to 95% and 90% confidence.

Confidence intervals are reported to indicate the degree of precision of our estimates. The narrower the confidence interval, the more precise the estimate. Because a small range of values contains the true parameter with high probability.

With the help of the CLT, it is not hard to find the CI for the sample mean $\bar{X}_n$. Let's set $\alpha = 5\%$, that is, we are trying to find the CI that contains the true mean 95% of the times. Assume our sample size n is large enough to invoke the CLT, we thus have

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

We want to find L and U such that

$$P(L \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq U) = 1 - 2\Phi(L) = 0.95$$

since the normal distribution is symmetric, $L = -U$. By looking at the CDF of standard normal, we get $L = -1.96$, $U = 1.96$. So the interval is

$$-1.96 \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq 1.96$$

With a little rearrangement, we have

$$\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}$$

Thus, the interval $\left[\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right]$ contains the true mean 95% of the times.

Theorem 7.8. *The* $100(1-\alpha)\%$ *confidence interval for the sample mean* $\bar{X}_n$ *is* $\bar{X}_n \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$*, where* $z_{\alpha/2}$ *is the critical value such that* $\Phi(z_{\alpha/2}) = \frac{\alpha}{2}$*.*

In practice, because we do not know $\sigma/\sqrt{n}$, we replace it with the standard error $s/\sqrt{n}$. Thus, we compute the confidence interval as $\bar{X}_n \pm z_{\alpha/2}SE$. However, this replacement is not without risk. When the sample size is small, s is a very poor estimate of σ. For the approximation to be valid, we require either the sample size is large enough ($n \geq 30$ at least) or the population distribution is nearly normal. Some commonly used confidence levels:

- 90% CI: $\alpha = 0.1$, $z_{0.05} = 1.645$
- 95% CI: $\alpha = 0.05$, $z_{0.025} = 1.96$
- 99% CI: $\alpha = 0.01$, $z_{0.005} = 2.58$

We go through some common misunderstandings about confidence intervals through an example.

Example 7.4. A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean

of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample. [2]

Solution: We know $\bar{X} = 3.2$, $s = 1.74$. The standard error is

$$SE = \frac{1.74}{\sqrt{50}} \approx 0.5$$

The approximate 95% CI is therefore

$$\bar{X} \pm 1.96SE \approx 3.2 \pm 0.5 = (2.7, 3.7).$$

Now check the following interpretations (true or false):

1. We are 95% confident that the average number of exclusive relationships in this sample is between 2.7 and 3.7.

   False. The CI definitely contains the sample mean $\bar{X}$.

2. 95% of college students have been in 2.7 to 3.7 exclusive relationships.

   False. The CI is about covering the population mean, not for covering 95% of the entire population.

3. There is 0.95 probability that the true mean falls in the interval (2.7, 3.7).

   False. The true mean μ is a fixed number, not a random one that happens with a probability.

4. The interval (2.7, 3.7) has probability of 0.95 of enclosing the true mean number of exclusive relationships of college students.

   False. The true mean is either in the interval or not. There is no uncertainty involved.

5. If a new random sample of size 50 is taken, we are 95% confident that the new sample mean will be between 2.7 and 3.7.

   False. The confidence interval is for covering the population mean, not for covering the mean of another sample.

6. This confidence interval is not valid since the number of exclusive relationships is integer-valued. Neither the population nor sample is normally distributed.

[2] This section and the next are based on Dr. Yibi Huang's lecture slides.

False. The construction of the CI only uses the normality of the sampling distribution of the sample mean (by the CLT). Neither the population nor the sample is required to be normally distributed.

So what is exactly the thing that has a 95% change to happen? It is the procedure to construct the 95% interval. About 95% of the intervals constructed following the procedure will cover the true population mean μ. After taking the sample and an interval is constructed, the constructed interval either covers μ or it doesn't. But if we were able to take many such samples and reconstruct the interval many times, 95% of the intervals will contain the true mean.

7.6 Hypothesis testing*

One day I woke up in the morning and came up with a question: Am I the average height of a Chinese man? I hypothesize that I am. My height is $\sqrt{3} \approx 1.73$ meters. Let μ be the average height of Chinese men. My hypothesis is

$$H_0 : \mu = \sqrt{3}$$

This is called the **null hypothesis**. I also suspect that the average Chinese men is taller than me (if the original hypothesis is false).

$$H_1 : \mu > \sqrt{3}$$

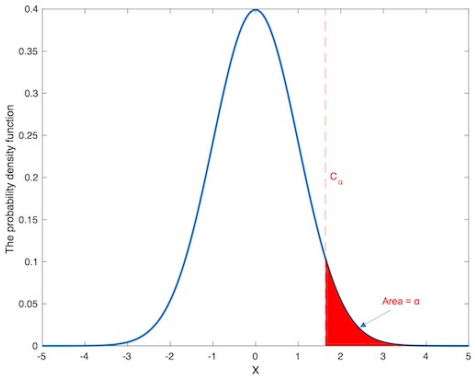
This is called the **alternative hypothesis**. How am I able to test which hypothesis is true? I can answer this question by collecting a small sample. Suppose I asked 50 people around me, and computed a sample average of $\bar{X} = 1.76$. Also assume we know the standard deviation $\sigma = 0.1$ (despite this is unrealistic). Does it prove or disprove the hypothesis?

By the CLT, we know $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is approximately standard normal. Suppose H_0 is true, we compute

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{1.76 - \sqrt{3}}{0.1/\sqrt{50}} = 1.98$$

which is know as the z-**score** of the sample mean. Thus, the probability of $\bar{X} > 1.76$ is

$$P(\bar{X} > 1.76) = 1 - \Phi\left(Z = 1.98\right) = 0.023$$



That is the small red region in the graph. That means, if H_0 is true, we only have a very small chance of observing $\bar{X} = 1.76$. Therefore, the hypothesis H_0 is likely false. In other words, if we reject H_0 and accept the alternative H_1, the probability that we have made a mistake is very low.

However, if we had observed $\bar{X} = 1.73$. The probability above is 0.56. That means we are very likely to observe this value if H_0 is true. In this case, it would be reasonable to accept H_0. In other words, we do not have strong enough evidence to reject the hypothesis.

The probability represented by the red area is called the p-**value**. The p-value is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct. A very small p-value means that such an extreme observed outcome would be very unlikely under the null hypothesis. Thus, The smaller the p-value, the stronger the evidence against the H_0.

In some studies, we can simply report the p-value and let people judge whether the evidence is strong enough. In other studies, we prefer to select a cut-off value α, call the **significance level**, and follow the rule:

- If the p-value $< \alpha$, reject H_0;
- If the p-value $> \alpha$, do not reject H_0.

Commonly used significance levels: 0.05 and 0.01. And we like to use the word "significant" to describe the test result:

- A test with p-value < 0.05 is said to be (statistically) **significant**;
- A test with p-value < 0.01 is said to be highly **significant**.

When we make a decision about accepting or rejecting a hypothesis, there are chances that we make a mistake. There are two types of mistakes: **Type 1 error** and **Type 2 error**.

		Decision	
		Reject H_0	Fail to reject H_0
Truth	H_0 is true	Type 1 error	✓
	H_0 is false	✓	Type 2 error

Type 1 error is rejecting the H_0 when it is true. **Type 2 error** is failing to reject the H_0 when it is false. Usually, it is more important to control the Type 1 error than the the Type 2 error. That is, we want to minimize the chance of falsely rejecting the null hypothesis.

In the example above, we reject the null hypothesis on the ground that there is only 2.3% of the chance that we could observe this sample. Therefore, the probability of Type 1 error is only 2.3%.

If we make decisions based on a significance level, the significance level is the Type 1 error rate. In other words, when using a 5% significance level, there is 5% chance of making a Type 1 error.

$$P(\text{Type 1 error}|H_0 \text{ is true}) = \alpha$$

This is why we prefer small values of α—smaller α reduces the Type 1 error rate. However, significance level doesn't control Type 2 error rate.

Hypothesis testing with z-statistics

We may have noticed that, in the above example, the assumption that the population σ is known is unrealistic. In practice, we approximate it with the

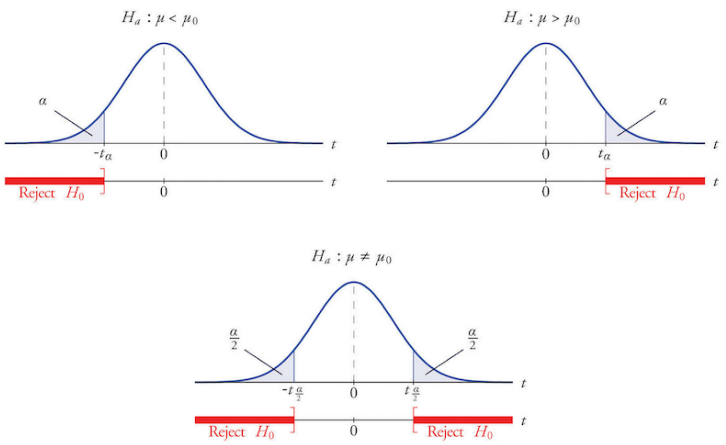
standard error $s/\sqrt{n}$. The approximate is valid if the the sample size is large enough or the underlying distribution is nearly normal. If this is not the case, we would opt for a t-test. Here we summarize the steps of testing for a population mean with z-statistics.

1. Set up the hypothesis:
   - $H_0 : \mu = \mu_0$
   - $H_1 : \mu <$ or $>$ or $\neq \mu_0$
2. Check assumptions and conditions
   - independent and identically distributed ($i.i.d$)
   - Nearly normal distribution or the sample size is large enough
3. Compute the test statistic and the p-value:

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

4. Make decision:
   - If p-value $< \alpha$, reject H_0
   - If p-value $> \alpha$, do not reject H_0

We notice that the **two-sided** hypothesis tests are very closed related to the concept of confidence intervals. A two-sided test means we are interested in rejection regions on both sides of the tail distribution. Typically, the alternative hypothesis is $H_1 : \mu \neq \mu_0$.



Suppose we are doing a hypothesis test under the significance level α, the region of accepting the H_0 is

$$-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{SE} \leq z_{\alpha/2}$$

such that the rejection region (p-value) has probability α. This is equivalent to

$$\bar{X} - z_{\alpha/2}SE \leq \mu \leq \bar{X} + z_{\alpha/2}SE$$

which is exactly the $100(1-\alpha)\%$ confidence interval of $\bar{X}$. Therefore, for a two-sided test, we have the rule:

- Reject H_0 if μ is not in the $100(1-\alpha)\%$ CI: $\bar{X} \pm z_{\alpha/2}SE$

We conclude this chapter by reiterating a couple of critical points that could be easily misunderstood.

Rejecting H_0 doesn't means we are 100% sure that H_0 is false. We might make Type 1 errors. Setting a significance level just guarantee we won't make Type 1 error too often.

Failing to reject H_0 does not necessarily mean H_0 is true. We could make a type 2 error when failing to reject H_0. Moreover, unlike type 1 error rate is controlled at a low level, type 2 error rate is usually quite high. When we fail to reject H_0, it just means the data are not able to distinguish between H_0 and

H_1. That's why we say *fail to reject*. p-value is not the probability that the H_0 is true.

Saying that results are statistically significant just informs the reader that the findings are unlikely due to chance alone. However, it says nothing about the practical importance of the finding. For example, rejecting the H_0: $\mu = \mu_0$ does not tell us how big the difference $|\mu - \mu_0|$ is. Mostly in practice we care more about the magnitude of this difference, rather than the fact that they are indeed different. It is possible that the difference is too small to be relevant even if it is significant.

Hypothesis testing with t-statistics

When the sample size is small, we opt for t-test for more reliable hypothsis testing. Define test statistics

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

where s is the sample standard deviation. For small samples, this test statistics follows a Student t-distribution with n degrees of freedom, $T \sim t(n)$.

Why Student-t distribution? Recall the definition of Student-t distribution: when the underlying distribution of $X_1, X_2, \ldots, X_n$ is Normal, sample variance s^2 follows a χ^2 distribution. T follows t distribution by definition regardless of the sample size. However, if the underlying distribution is not normal, this argument loses ground. We use t-test mainly as a convention. But t distribution has heavier tails than standard normal, meaning that we are more likely to reject a hypothesis based on t distribution. In other words, t-test is a more conservative choice than z-test for small samples.

one-tail α	0.05	0.025	0.005
two-tail α	0.10	0.05	0.01
d.f.			
10	1.812	2.228	3.169
20	1.725	2.086	2.845
30	1.697	2.042	2.750
z value	1.645	1.960	2.576

The table shows a few critical values for t-test with different degrees of freedom (d.f.). We can see as the sample size gets larger, t distribution converges to standard normal.