Probability and Statistics

Naz Koont

21 July 2021

Overview

1. Probability
   - Probability Space
   - Random Variable
   - Probability Distributions: CDF and PDF
   - Moments: Expectation, variance, mgf
   - Conditional and Joint Probability, Independence
   - Bayes Rule
2. Statistics
   - Estimators
   - LLN and CLT
   - Hypothesis Testing, p-values
   - Confidence Intervals
3. Econometrics
   - CEF and OLS
   - Timeseries data (stationarity)

Motivation

Financial economists use empirical analyses to test the validity of economic models and estimate relationships between variables. The framework and foundation for statistical analysis comes from probability theory.

> The most important questions of life are, for the most part, really only problems of probability.
> – Pierre-Simon Laplace

Probability Spaces

We have already developed the tools that we need to formalize the notion of probability. Let us consider the following triplet of mathematical objects:

Definition (Probability Space)
A probability space $(\Omega, \mathcal{A}, \mathbb{P})$ consists of three elements:

1. A **sample space** Ω, which is the set of all possible outcomes.
2. An **event space** $\mathcal{A}$, which a set of events i.e. sets of outcomes in the sample space.
3. A **probability measure** (function) $\mathbb{P}$, which assigns each event in the event space $\mathcal{A}$ a probability, $\mathbb{P} : \mathcal{A} \to [0, 1]$.

Probability Measure

Theorem (Properties of $\mathbb{P}$)

Probability measure $\mathbb{P}$ must satisfy the following properties:

1. *Positivity:* $\mathbb{P}(A) \geq 0, \forall A \in \mathcal{A}$
2. *Normalization:* $\mathbb{P}(\Omega) = 1$
3. *Additivity: If* $A \cap B = \emptyset$, *then* $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ *and in general,* $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$
4. *Monotonicity: If* $A \subseteq B$, *then* $\mathbb{P}(A) \leq \mathbb{P}(B)$

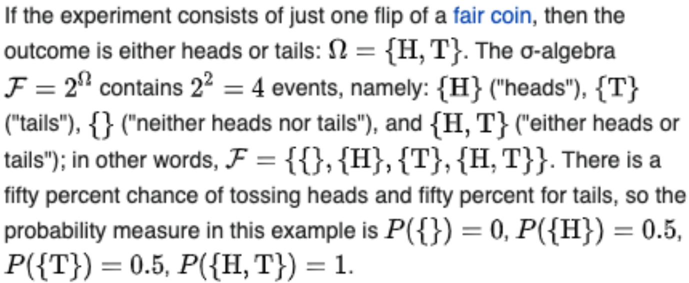
Useful Concepts

Lets further define two concepts that are important going forward:

Definition (σ-Algebra)
If the nonempty collection $\mathcal{A}$ of subsets of $\Omega \neq \emptyset$ contains all complements and unions of its sets (i.e. is "closed" under complements and countable unions), then it is a σ-algebra of Ω.

Definition (Measurable Space)
The pair $(\Omega, \mathcal{A})$, where $\mathcal{A}$ is a σ-algebra of subsets of Ω is called a measurable space. The subsets of $\mathcal{A}$ are called the events of Ω.

Probability Space: Example



If the experiment consists of just one flip of a fair coin, then the outcome is either heads or tails: $\Omega = \{\mathrm{H}, \mathrm{T}\}$. The σ-algebra $\mathcal{F} = 2^{\Omega}$ contains $2^2 = 4$ events, namely: $\{\mathrm{H}\}$ ("heads"), $\{\mathrm{T}\}$ ("tails"), $\{\}$ ("neither heads nor tails"), and $\{\mathrm{H}, \mathrm{T}\}$ ("either heads or tails"); in other words, $\mathcal{F} = \{\{\}, \{\mathrm{H}\}, \{\mathrm{T}\}, \{\mathrm{H}, \mathrm{T}\}\}$. There is a fifty percent chance of tossing heads and fifty percent for tails, so the probability measure in this example is $P(\{\}) = 0$, $P(\{\mathrm{H}\}) = 0.5$, $P(\{\mathrm{T}\}) = 0.5$, $P(\{\mathrm{H}, \mathrm{T}\}) = 1$.

Random Variables

Along with a probability measure, a central concept in probability and statistics is that of a *random variable*. Informally, this is a function whose values depend on the outcomes of random phenomena.

Definition (Random Variable)
Let $(\Omega, \mathcal{A})$ and $(\mathbb{R}, \mathcal{B})$ be two measurable spaces. The map $X : \Omega \to \mathbb{R}$ is called real-valued random variable, if for all $B \in \mathcal{B}$, $X^{-1}(A)$ is **measurable** for all $B \in \mathcal{B}$, i.e.:

$$X^{-1}(B) = \{\omega \in \Omega | X(\omega) \in B\} \in \mathcal{A}$$

Thus by definition a random variable takes events from one sample space into well-defined events in another space.

Probability Distribution

If a random variable $X : \Omega \to \mathbb{R}$ defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is given, we can ask questions like, "How likely is it that the value of X is equal to 2?".

- This is the same as the probability of the event $\{\omega | X(\omega) = 2\} \equiv \mathbb{P}(X = 2) = \mathbb{P}_X(2)$.
- In other words, since $\mathbb{P}$ specifies the probability of every subset of Ω, it also induces probabilities on events expressed in terms of X.
- Recording all of these probabilities of output ranges of a real-valued random variable X yields the **probability distribution** of X.
- The probability distribution "forgets" about the particular probability space used to define X and only records the probabilities of various values of X.

Cumulative Distribution Function (CDF)

The probability distribution "forgets" about the particular probability space used to define X and only records the probabilities of various values of X. Such a probability distribution can always be captured by its **cumulative distribution function**.

Definition (Cumulative Distribution Function)
The cumulative distribution function (c.d.f) of a real-valued random variable $X : \Omega \to \mathbb{R}$ is defined as the real-valued function $F : \mathbb{R} \to [0, 1]$ given by:

$$F_X(x) = \mathbb{P}\{\omega \in \Omega | X(\omega) \leq x\} = \mathbb{P}\{X^{-1}(-\infty, x]\}$$

Thus it is simply its induced probability measure $\mathbb{P}_X$ evaluated at sets of the form $(-\infty, x]$.

Properties of the CDF

Theorem (Properties of the CDF)

If F_x is the CDF of a random variable $X : \Omega \to \mathbb{R}$, then

1. *F_X is non-decreasing.*
2. $\lim_{x\uparrow\infty} F_X(x) = 1$
3. $\lim_{x\downarrow-\infty} F_X(x) = 0$
4. $\lim_{h\to0^+} F_X(x+h) = F_X(x)$ *i.e. the CDF is right-continuous.*

PDF and PMF

Definition (Probability Density Function)
For continuous random variables, assuming that F_X is differentiable, by the fundamental theorem of calculus we have that:

$$F_X(x) = \int_{-\infty}^{x} p_x(x)dx$$

The function $p_x(x)$ is called the probability density function (PDF) of X. Note that $P_X(x) = \mathbb{P}(X = x)$.

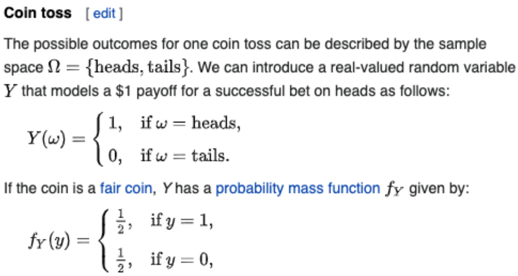
Definition (Probability Mass Function)
Scalar random variables that take values in a discrete set such as $\{1, 2, ...\}$ do not have a density function. Instead they have a probability mass function (PMF), with the same interpretation as the density.

$$p_X(x) = \sum_i \mathbb{P}(X = x_i)1\{x = x_i\}$$

The function $p_X(x)$ is called the probability mass function (PMF) of X. Note that $P_X(x) = \mathbb{P}(X = x)$.

Examples of CDFs, PDFs: Bernoulli

Bernoulli with $p = .5$.



Coin toss [edit]

The possible outcomes for one coin toss can be described by the sample space $\Omega = \{\text{heads}, \text{tails}\}$. We can introduce a real-valued random variable Y that models a \$1 payoff for a successful bet on heads as follows:

$$Y(\omega) = \begin{cases} 1, & \text{if } \omega = \text{heads}, \\ 0, & \text{if } \omega = \text{tails}. \end{cases}$$

If the coin is a fair coin, Y has a probability mass function f_Y given by:

$$f_Y(y) = \begin{cases} \frac{1}{2}, & \text{if } y = 1, \\ \frac{1}{2}, & \text{if } y = 0, \end{cases}$$

Definition (Bernoulli Distribution)

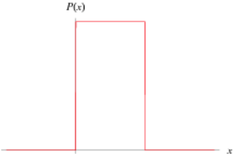
The random variable X is said to have a Bernoulli distribution with parameter $p \in [0, 1]$ if the CDF is given by

$$F_X(x) = \begin{cases} 0 & x \leq 0 \\ 1 - p & x \in [0, 1) \\ 1 & x \geq 1 \end{cases}$$

Examples of CDFs, PDFs: Uniform

Definition (Uniform Distribution)
The random variable X is said to have a uniform distribution in $[a, b]$ if the CDF is given by

$$F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a, b] \\ 1 & x > b \end{cases}$$



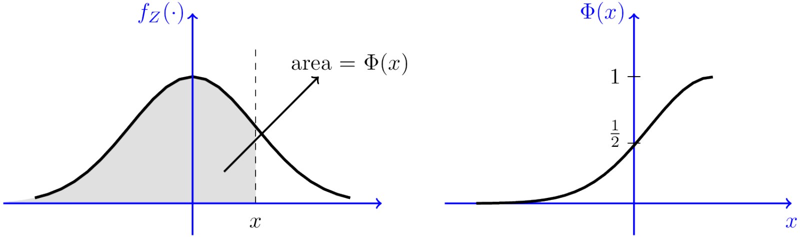
The PDF is given of X by

$$f_X(x) = \begin{cases} 0 & x < a \\ \frac{1}{b-a} & x \in [a, b] \\ 0 & x > b \end{cases}$$

Examples of CDFs, PDFs: Normal

Definition (Gaussian (Normal) Distribution)
The random variable X is said to have a Normal distribution with parameters (μ, σ^2) if the CDF is given by

$$\Phi(x) = F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(u-\mu)^2\right) du$$



Note that a Bernoulli random variable is discrete, whereas Uniform and Normal random variables have continuous support.

Expectation of $X : \Omega \to \mathbb{R}$

Let us introduce the definition of expectation for discrete and (absolutely) continuous random variables.

Definition (Expectation of continuous random variable)
Let $X : \Omega \to \mathbb{R}$ an absolutely continuous random variable with CDF F and PDF $f(x)$. Let $g : X \to \mathbb{R}$ a real-valued function. The expectation of $g(X)$ under F is defined as:

$$\mathbb{E}_F[g(X)] = \int_{\mathbb{R}} g(x)f(x)dx$$

Definition (Expectation of discrete random variable)
Let $\{x_1, x_2, ...\}$ the support of discrete random variable $X : \Omega \to \mathbb{R}$. Let $g : X \to \mathbb{R}$ a real-valued function. The expectation of $g(X)$ under PMF p is defined as:

$$\mathbb{E}_F[g(X)] = \sum_{x_i \in Supp} g(x_i)p(x_i)$$

Moments of $X : \Omega \to \mathbb{R}$

Lets calculate some expectations:

1. Let $X \sim Bernoulli(p)$.

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x_i \in Supp} xp(x_i) \\ &= \mathbb{P}(X = 1) * 1 + \mathbb{P}(X = 0) * 0 \\ &= p * 1 + (1 - p) * 0 = p\end{aligned}$$

2. Let $X \sim U(a, b)$.

$$\begin{aligned}\mathbb{E}_F[g(X)] &= \int_{\mathbb{R}} xf(x)dx \\ &= \int_{-\infty}^{a} 0dx + \int_{a}^{b} \frac{x}{b-a}dx + \int_{b}^{\infty} 0dx \\ &= \frac{1}{b-a}\frac{x^2}{2}|_a^b \\ &= \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}\end{aligned}$$

Moments of $X : \Omega \to \mathbb{R}$

Some important moments are the following:

Definition (Moments)

1. Mean of a RV: $\mathbb{E}_F[X] = \mu$ is called the **mean** or **first moment** of X.
2. Variance of RV: $\mathbb{E}_F[(X - \mu)^2] = \sigma^2$ is called the **variance** or **second centered moment** of X.
3. k-th moment of a RV: $\mathbb{E}_F[X^k] = \mu$ is called the **k-th uncentered moment** of X.
4. Probability: $\mathbb{P}(X \in \mathcal{A}) = \mathbb{E}[1\{X \in \mathcal{A}\}]$

Important properties of moments:

Theorem (Properties of Moments)

1. *Jensen's Inequality: For any convex function g:* $\mathbb{E}_F[g(X)] \geq g(\mathbb{E}[X])$
2. *Markov's Inequality: For any random variable X and $c \geq 0$,* $c\mathbb{P}_X[|X| > c] \leq \mathbb{E}_F[|X|]$

Moments of $X : \Omega \to \mathbb{R}$

Theorem (Properties of Expectation)

1. *Expectation is linear:* $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$ *and* $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
2. *For independent random variables,* $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$

Theorem (Properties of Variance)

1. *Variance of* X*:* $Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
2. *Covariance between* X, Y*:* $Cov(X, Y) = \mathbb{E}[(X - \mu_x)(Y - \mu_y)] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
3. $Var(aX + b) = a^2 var(X)$
4. *For independent* X, Y*:* $Var(X + Y) = Var(X) + Var(Y)$
5. *In general,* $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$

Moments Generating Functions

Definition (Moment Generating Function*)
The random variable X is said to have a moment generating function (MGF) $m_x(t)$ if

$$m_x(t) \equiv \mathbb{E}_F[exp(tX)] \leq \infty \quad \text{for all } t \in (-\epsilon, \epsilon) \text{ for some } \epsilon > 0$$

- Two random variables with the same moment generating function have the same distribution.
- As the name implies, the MGF can be used to compute a distribution's moments: the nth moment about 0 is the nth derivative of the MGF, evaluated at 0.

Dependence between random variables

If we could only analyze one random variable at a time, the questions that we would be able to answer would not be very interesting. Most of the empirical questions in economics boil down to *what is the causal effect of X on Y?*.

- So, we need a way to describe the relationship between at least two random variables.
- Until now, we have been considering probability density and mass functions assuming that we have *no* information about the random variable of interest.
- We might be able to do better if we observe something that is informative about that variable, and incorporate that information into our expectation.

Definition (Conditional Probability)
Consider two events $A, B \in \mathcal{A}$. The conditional probability that A occurs given that B occurs is:

$$\mathbb{P}(A|B) \equiv \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Bayes Rule

Now we come to a very important result for applied applications. Often we know a conditional probability in one direction, and would like to know the conditional probability in the other direction.

Theorem (Bayes Rule)

Recall the conditional probability $\mathbb{P}(A|B)$. Bayes Rule is a formula for the "inverse" conditional probability, $\mathbb{P}(B|A)$:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}$$

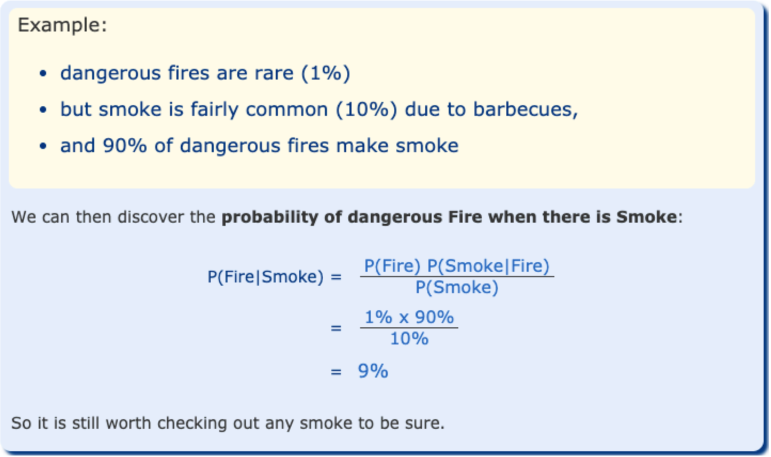
Definition (Independence)
Two events A and B are said to be independent if $\mathbb{P}(A|B) = \mathbb{P}(A)$. In other words, B provides no information about whether A has occurred.

Bayes Rule

Lets derive Bayes Rule from the definition of conditional probability and some probability rules:

$$
\begin{aligned}
\mathbb{P}(A|B) &\equiv \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \\
\mathbb{P}(B|A) &\equiv \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \\
\implies \mathbb{P}(A \cap B) &= \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A) \\
\mathbb{P}(A|B) &= \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}
\end{aligned}
$$

Bayes Rule



Example:

- dangerous fires are rare (1%)
- but smoke is fairly common (10%) due to barbecues,
- and 90% of dangerous fires make smoke

We can then discover the **probability of dangerous Fire when there is Smoke**:

$$\begin{aligned} P(\text{Fire}|\text{Smoke}) &= \frac{P(\text{Fire})\ P(\text{Smoke}|\text{Fire})}{P(\text{Smoke})} \\ &= \frac{1\% \text{ x } 90\%}{10\%} \\ &= 9\% \end{aligned}$$

So it is still worth checking out any smoke to be sure.

Joint CDF

Definition (Multivariate or Joint CDF)
To fully understand the relationship between two random variables X, Y, we need to look at something that fully characterizes their joint distribution, such as their joint CDF.

$$F_{X,Y}(x, y) = \mathbf{P}\left[(X \leq x) \cap (Y \leq y)\right]$$

From this joint distribution, we can recover the marginal univariate PDF of X for example by integrating out Y:

$$f_x(x) = \int_{S_Y} f_{X,Y}(x, y)dy$$

Theorem (Bayes Rule For Joint PDF)

The Bayes Rule applied to a joint pdf becomes:

$$f_{X|Y}(x; y) = \frac{f_{Y|X}(y; x)f_y(y)}{f_X(x)}$$

Statistics

By a small sample, we may judge the whole piece.
– Miguel De Cervantes, Don Quixote

Many statistical inference problems can be identified as being one of three types: estimation, confidence sets, or hypothesis testing. Here I will give a brief introduction to these ideas. To learn more about these topics, take the Econometrics courses offered at CBS.

Basics of Estimators

The frequentist approach to estimation starts with the premise that there is a *population parameter* θ that determines how we observe data.

- We observe a *sample* of this data, $\{x_i\}_{i=1}^N = X$ and use this sample to produce an *estimate* of θ, the property of the population that we would like to know about.
- Our prime objective is to estimate properties of the population, using a sample and an estimator.

Definition (Estimator)
For a random variable X, an estimator $\hat{\theta} : X \to \mathbb{R}^n$ is a function that maps the sample space to a set of sample estimates.

- The estimator itself is a random variable.
- The estimate for a given realization of the data is a fixed number.

A **point estimator** outputs a single number regarded as the most plausible value of θ. An **interval estimator** outputs a range of numbers. This range regarded as likely containing the true value of θ.

Sample Mean

An important example of a point estimator is the sample mean, which is useful to arrive at point estimates of a given parameter.

Definition (Sample Mean)
The sample mean is the average of the values of a random variable in a sample, used as an estimator for the population mean of a random variable, μ.

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

It turns out that this estimator is a good one for the population mean, as it gets very close to the population mean as the sample size grows.

Theorem (Law of Large Numbers)

The sample mean $\bar{x}$ converges in probability towards μ as $n \to \infty$.

$$\lim_{n \to \infty} \mathbb{P}(|\bar{x}_n - \mu| > \epsilon) = 0 \textit{ for all } \epsilon > 0$$

(Convergence in probability vs convergence in distribution)

Standard Error

Given a point estimator $\hat{\theta}$, we can define the standard error of the estimator by recalling that the estimator is itself a random variable:

Definition (Standard Error)
The standard error of an estimator $\hat{\theta}$ is its standard deviation:

$$SE(\hat{\theta}) = \sqrt{Var(\hat{\theta})}$$

As an example, we can calculate the standard error of the sample mean:

$$SE(\bar{x}) = \sqrt{var(\bar{x})} = \sqrt{\frac{1}{n^2}\sum_i var(x_i)} = \sqrt{\frac{1}{n^2}n\sigma} = \frac{\sigma}{\sqrt{n}}$$

Sampling Distribution & Central Limit Theorem

Since an estimator is itself a random variable, it has a distribution as well, called the *sampling distribution* of an estimator $\hat{\theta}$.

However, the following important theorem tells us that we do not need to worry too much about the specific sampling distribution of an estimator under certain regularity conditions, because the shape of the sampling distribution will approach normality as the sample size N increases:

Theorem (Central Limit Theorem)

If $X_1, ...X_n$ are n random samples drawn from a population with overall mean μ and finite variance σ^2, and if $\bar{X}$ is the sample mean, then the limiting form of the following distribution is $N(0,1)$:

$$Z = \lim_{n \to \infty} \sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sigma} \right) \sim N(0,1)$$

Sampling Distribution & Central Limit Theorem

Theorem (Central Limit Theorem)

If $X_1, ...X_n$ are n random samples drawn from a population with overall mean μ and finite variance σ^2, and if $\bar{X}$ is the sample mean, then the limiting form of the following distribution is $N(0,1)$:

$$Z = \lim_{n \to \infty} \sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sigma} \right) \sim N(0,1)$$

This theorem is useful in economic regression analysis as the assumption of normally distributed errors simplifies much analysis and can be justified by the central limit theorem– regarding the error as the sum of many independent errors. A similar logic can be used to explain why so many variables in real-world data sets seem to be normally distributed (although not always in financial data– see the Time Series Analysis section below).

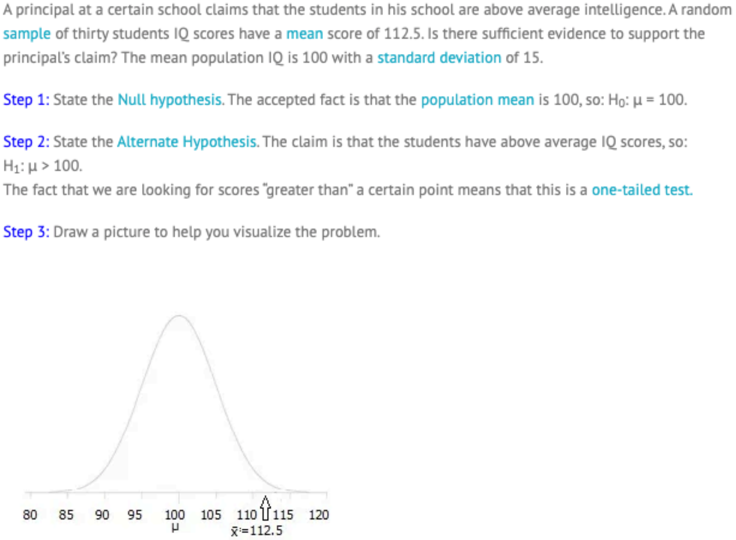
Hypothesis Testing

In Hypothesis testing, we start with some default theory about a population parameter called the **null hypothesis**, and ask if the data provide sufficient evidence to reject the theory, or if we fail to reject the null hypothesis.

- For example, we may state a null hypothesis $H_0 : \theta = 0.5$ and an alternative hypothesis $H_A : \theta \neq 0.5$.
- Then we use our **test statistic**, such as the sample mean, in order to learn about θ.
- And finally, we need a **decision rule** about rejecting H_0 based on how far $t = \bar{x} - 0.5$ is from 0.
- Since large $|t|$ would be evidence against H_0, a rule could be to reject H_0 if and only if $|t| > t_c$ for some **critical value** t_c.
- This threshold value increases as we wish to be more confident that we are not falsely rejecting H_0.

Hypothesis Testing

For example, let's look at a z-test (distribution of test statistic approximated by normal distribution).



A principal at a certain school claims that the students in his school are above average intelligence. A random sample of thirty students IQ scores have a mean score of 112.5. Is there sufficient evidence to support the principal's claim? The mean population IQ is 100 with a standard deviation of 15.

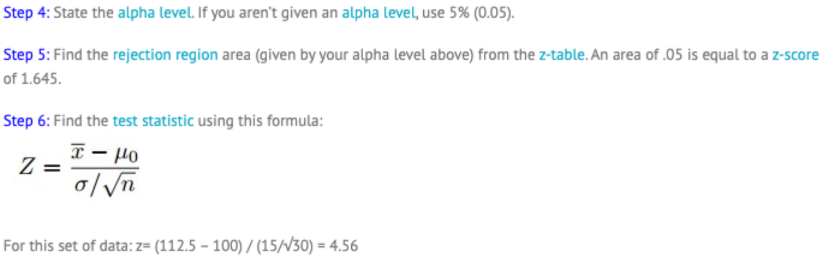
Step 1: State the Null hypothesis. The accepted fact is that the population mean is 100, so: H_0: μ = 100.

Step 2: State the Alternate Hypothesis. The claim is that the students have above average IQ scores, so: H_1: μ > 100.
The fact that we are looking for scores "greater than" a certain point means that this is a one-tailed test.

Step 3: Draw a picture to help you visualize the problem.

Hypothesis Testing

Next we calculate the distance between sample mean and population mean in units of the standard error $\sigma/\sqrt{n}$ (z-score).



Step 4: State the alpha level. If you aren't given an alpha level, use 5% (0.05).

Step 5: Find the rejection region area (given by your alpha level above) from the z-table. An area of .05 is equal to a z-score of 1.645.

Step 6: Find the test statistic using this formula:

$$Z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

For this set of data: z= (112.5 – 100) / (15/√30) = 4.56

If Step 6 is greater than Step 5, reject the null hypothesis. If its less than Step 5, you cannot reject the null hypothesis. In this case, it is more ($4.56 > 1.645$), so you can reject the null.

Hypothesis Testing and p-values

Another popular way of reporting statistical significance is with a ***p*-value**.

- The *p*-value associated with a hypothesis test is defined as the probability of observing a test statistic at least as extreme as the one we actually observed, assuming that H_0 is true.
- The benefit of reporting a *p*-value is that it allows the reader to test your hypothesis at their choice of confidence α, rather than the one you selected.

Confidence Intervals

Now we can give an important example of an interval estimator, a **confidence interval**. This reports all of the values of θ_0 for which we would fail to reject H_0 in the test $H_0 : \theta = \theta_0$ and $H_A : \theta \neq \theta_0$. This is useful because it reports, holding fixed the confidence α, **all** of the null hypotheses that would not be rejected.

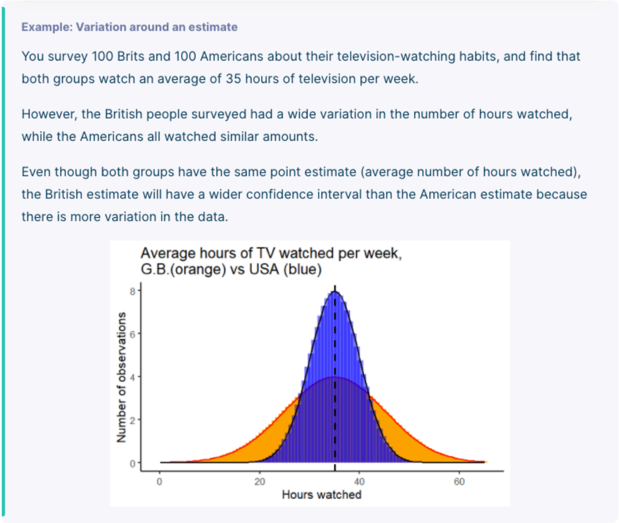
Definition (Confidence Interval)

A confidence interval $C(X)$ for a parameter θ is a set of possible values which contains θ with some specified probability.

A note of warning here – the confidence interval is not a probability statement about θ, which is a fixed quantity, not a random variable.

- One interpretation is that if the same experiment is repeated over and over again, the constructed interval will contain the parameter 95% of time.
- A higher confidence interval generates a wider (less precise) confidence interval. Often 95% is used.
- In contrast, Bayesian inference methods (rather than Frequentist inference) treat θ as a random variable, and do make probability statements about θ.

Confidence Intervals



Example: Variation around an estimate

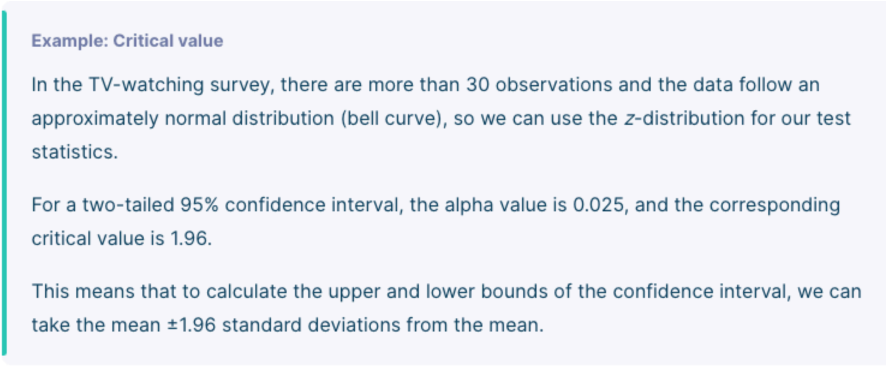
You survey 100 Brits and 100 Americans about their television-watching habits, and find that both groups watch an average of 35 hours of television per week.

However, the British people surveyed had a wide variation in the number of hours watched, while the Americans all watched similar amounts.

Even though both groups have the same point estimate (average number of hours watched), the British estimate will have a wider confidence interval than the American estimate because there is more variation in the data.

Average hours of TV watched per week,
G.B.(orange) vs USA (blue)
Number of observations
0
2
4
6
8
0
20
40
60
Hours watched

Confidence Intervals



Example: Critical value

In the TV-watching survey, there are more than 30 observations and the data follow an approximately normal distribution (bell curve), so we can use the *z*-distribution for our test statistics.

For a two-tailed 95% confidence interval, the alpha value is 0.025, and the corresponding critical value is 1.96.

This means that to calculate the upper and lower bounds of the confidence interval, we can take the mean ±1.96 standard deviations from the mean.

CEF And Ordinary Least Squares

We are interested in explaining economic relationships, i.e. in explaining "systematic randomness". Thus, the object that we would like to estimate is the population relationship between two random variables X and Y. This relationship is conveniently summarized by the *Conditional Expectation Function*.

Definition (Conditional Expectation)
The conditional expectation for a "dependent" variable Y_i, given a set of covariates X_i, is the expectation, or population average, of Y_i, with X_i held fixed.

$$E[Y_i|X_i = x]$$

The conditional expectation function is useful because it is the "best" predictor of Y_i given X_i, in the sense that it solves a minimum mean squared error (MMSE) prediction problem.

CEF And Ordinary Least Squares

Theorem (The CEF Prediction Property)

$$E[Y_i|X_i] = \arg\min_{m(X_i)} \mathbb{E}[(Y_i - m(X_i))^2]$$

There is a tight link between the regression function and the CEF. The regression function is the “best” linear approximation to the CEF, the “best” linear predictor of Y_i given X_i in the sense that it solves a MMSE prediction problem, and is equivalent to the CEF if the CEF is linear. *Therefore, one should be interested in regression parameters if one is interested in the CEF.*

CEF And Ordinary Least Squares

To arrive at the population regression function, we are seeking a vector of population regression coefficients, defined as a solution to a population least squares problem. Let the $K \times 1$ regression coefficient vector be defined by solving:

$$\beta = \arg\min_{b} \mathbb{E}[(Y_i - X_i' b))^2]$$

To optimize this function, we find the first order condition:

$$\mathbb{E}[X_i(Y_i - X_i' b)] = 0$$

And thus the solution can be written:

$$\beta = \mathbb{E}[X_i X_i']^{-1} \mathbb{E}[X_i Y_i] \qquad (\textit{invertibility})$$

And so the slope coefficient for the kth regressor x_k is:

$$\beta_k = \frac{Cov(Y_i, \tilde{x}_{ik})}{Var(\tilde{x}_{ik})}$$

Where $\tilde{x}_{ik}$ is the residual of a regression of $x_{i,k}$ on all the other covariates, $x_{i,-k}$, i.e. $x_{i,k}$ after *partialling out* the effects of other regressors.

Standard Error of OLS β

We have some parameter which we estimate:

$$\beta_k = \frac{Cov(Y_i, \tilde{x}_{ik})}{Var(\tilde{x}_{ik})}$$

As we have discussed, this is an estimate of the true parameter β, and we are interested in understanding the distribution of estimate if we repeatedly drew random samples from the same population and estimated β. Thus, we want the variance of the sample average effect.

Definition (Standard Error)
The standard error of an estimator $\hat{\theta}$ is its standard deviation:

$$SE(\hat{\theta}) = \sqrt{Var(\hat{\theta})}$$

So in order to calculate the standard error of $\hat{\beta}$, first we would like to get the variance of $\hat{\beta}$

Standard Error of OLS β

$$\begin{aligned}
var(\hat{\beta}) &= var\left[\frac{cov(X,Y)}{var(Y)}\right] \\
&= var\left[\frac{\sum(X_i-\mu_X)(Y_i-\mu_Y)}{\sum(X_i-\mu_X)^2}\right] \\
&= \frac{1}{(\sum(X_i-\mu_X)^2)^2} var\left[\sum(X_i-\mu_X)\epsilon_i\right] \\
&= \frac{\sum(X_i-\mu_X)^2}{(\sum(X_i-\mu_X)^2)^2} var\left[\epsilon_i\right] \qquad (*) \\
SE(\hat{\beta})^2 &= \frac{\sigma_\epsilon^2}{\sum(X_i-\mu_X)^2}
\end{aligned}$$

Where we have assumed that there is some constant σ_ϵ^2 that does not vary with X. In contrast, we can instead derive *White* standard errors, or robust standard errors that do not require the homoskedasticity assumption.

Time Series Analysis

Time series data, such as stock market prices and returns, is very important in financial analysis.

- This data violates many of the common assumptions used in statistical analysis and require different techniques.
- For details, take the PhD Timeseries Analysis course offered at CBS.
- For one salient example, let's look at the non-stationarity of prices.

A basic property we would like to have of our data is that the data properties should not be changing too much over time, or in ways that imply there are fundamental differences between the data collected yesterday and today.

Stationarity intuitively refers to something that is not changing much over time. We usually apply this concept to moments in the data, such as means and variances.

Variance of Returns Scales with Time

Let us consider an example to see whether different types of financial data satisfy this property.

Given a process for returns with mean μ and standard deviation σ, consider the log value of a portfolio p_t that earns a return each period and reinvests all wealth.

$$p_t = p_{t-1} + r_t = p_{t-1} + \mu + \sigma\epsilon_t$$

This value process is said to follow a *Random Walk with Drift*, i.e. is a process with unforecastable increments, except for a constant drift term μ. Simulated data from this model approximates the properties of stock return data.

This is the original Efficient Markets model of Gene Fama (1970): if markets are efficient, returns are not forecastable apart from the constant risk premium component.

Variance of Returns Scales with Time

$$p_t = p_{t-1} + r_t = p_{t-1} + \mu + \sigma \epsilon_t$$

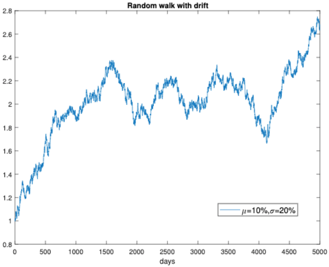
In this model, prices are *non-stationary* whereas returns are *stationary*, in the sense of the stability of their covariances over time. Consider the variance of p_t for $\mu = 0$:

$$\lim_{t \to \infty} Var[p_t | p_0] = \lim_{t \to \infty} t\sigma^2 = \infty$$

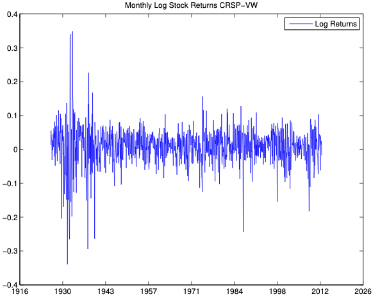
Instead, consider the return process $r_t = \mu + \sigma \epsilon_t$:

$$\lim_{t \to \infty} Var[r_t] = \lim_{t \to \infty} Var[\mu + \sigma \epsilon_t] = \sigma^2, \forall t$$

Non-stationary vs Stationary processes



A simulated Random Walk



Overview

1. Probability
    - Probability Space
    - Random Variable
    - Probability Distributions: CDF and PDF
    - Moments: Expectation, variance, mgf
    - Conditional and Joint Probability, Independence
    - Bayes Rule
2. Statistics
    - Estimators
    - LLN and CLT
    - Hypothesis Testing, p-values
    - Confidence Intervals
3. Econometrics
    - CEF and OLS
    - Timeseries data (stationarity)