DIW GRADUATE CENTER

# Introduction to Probability and Statistics

Anton Velinov

# Takeaways

- $\sigma$-algebra
- Probability
- Random variables
- Density functions
- Change of variable technique
- Expectation and population moments
- Classical estimation and estimator properties
- Convergence results

# Introduction I

- Say we can perform a repeatable controlled experiment, whose outcome we cannot predict with certainty
- Such an experiment is known as a *random experiment*
- Each conceivable outcome is a *sample point*
- The set of all possible outcomes is the *sample space*, $S$
- A subset of $S$ is referred to as an *event*.

# Introduction II

## Class Exercise

What is the sample space, $S$ of the experiment of rolling a fair die once?

## Class Exercise

What is the sample space, $S$ of the experiment of tossing a fair coin until a head appears?

## Definition

If $S$ contains a finite or countably infinite number of events it is said to be *discrete*. If the sample space contains an uncountably infinite number of elements (eg. an interval on $\mathbb{R}$), it is said to be *continuous*.

# $\sigma$-algebra I

## Definition: Power Set

The *Power set* of a set $S$, written as $\mathcal{P}(S)$ or $2^S$ is the set of all subsets of $S$. If $S$ is discrete with $n$ elements, then $\mathcal{P}(S)$ contains $2^n$ elements.

## Definition: $\sigma$-Algebra

Let $\Sigma \subset \mathcal{P}(S)$ such that the following hold:

- $\Sigma$ is non empty: $\Sigma \neq \emptyset$. There is at least one $D \subset S$ in $\Sigma$
- If $D$ is in $\Sigma$, then $D^c$ is also in $\Sigma$
- If $D_1, D_2, \ldots, D_n$ are in $\Sigma$, then the unions $D_1 \cup D_2 \cup D_3 \cdots D_n$ are also in $\Sigma$.

Then $\Sigma$ is known as a *$\sigma$-algebra* or a *$\sigma$-field*.

## Definition: Borel $\sigma$-algebra

In case $S$ is continuous, a similar concept known as a *Borel $\sigma$-algebra* is applied.

# $\sigma$-algebra II

## Intuition

We need to be able to assign some form of measure (probability) to events. A $\sigma$-algebra allows us to do just that. Elements of the $\sigma$-algebra are called *measurable sets* and the ordered pair $(S, \Sigma)$ is a *measurable space*.

## Class Exercise

Show that the above definition of a $\sigma$-algebra necessarily implies that both $S$ and $\emptyset$ are in $\Sigma$.

## Class Exercise

Suppose $S = \{a, b, c\}$

1. What is $\mathcal{P}(S)$?
2. Would $\{\emptyset, \{a\}, \{a, b\}, \{b, c\}, \{a, b, c\}\}$ constitute a $\sigma$-algebra? Explain.

# Probability, $P(\cdot)$

## Definition

- The ***classical*** interpretation concerns mutually exclusive and equally likely outcomes. Eg. the probability of $X = 1$ when tossing a fair die is 1/6. But what if the die is not fair?
- The ***relative frequency*** interpretation supposes an experiment can be replicated $N$ times. We then observe the number of times, $q$ the event occurs and the probability is defined as $\lim_{N\to\infty} q/N$. This is the most common definition used in statistics. But what if we can't exactly replicate an experiment?
- Sometimes a researcher may have a preconceived notion of the probability of an event, such ***subjective*** probability is often used in Bayesian analysis. But how can that be judged?

# Axioms of $P(\cdot)$

## Axioms of $P(\cdot)$

$P(\cdot)$ is a real valued function that assigns a real number to every member of the $\sigma$-algebra, ($P : \Sigma \to \mathbb{R}$) such that the following axioms are satisfied:

- For any event $A \subset S$ we have $0 \leq P(A) \leq 1$
- $P(S) = 1$
- For disjoint events $A_i, i = 1, 2, \ldots$ we have $P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \cdots$

## Class Exercise

Show the last axiom by means of a Venn diagram.

# Properties of $P(\cdot)$

## Properties of $P(\cdot)$

- $P(\emptyset) = 0$
- For any $A \subset S, P(A^c) = 1 - P(A)$
- If $A \subseteq B, P(A) \leq P(B)$
- $P(A \cup B) = P(A) + (B) - P(A \cap B)$.

## Class Exercise

Show the last property by means of a Venn diagram.

# Conditional Probability

## Definition

The *conditional probability* of event $A$ given that event $B$ will occur or has occurred is written as $P(A|B)$.

## Class Exercise

For a die let $A = \{1\}$ and $B = \{1, 3, 5\}$.

1. What is the unconditional probability $P(A)$?
2. What is the conditional probability $P(A|B)$?

In general

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

# Independent Events

## Definition

When knowing that event $B$ has/will occur does not affect the probability that event $A$ will occur, formally $P(A|B) = P(A)$, we then say that $A$ and $B$ are *independent events*.

Hence,

$$P(A \cap B) = P(A)P(B).$$

## Class Exercise

Consider the experiment of tossing a coin three times and define the events: $A = \{\text{Head on each of the first 2 tosses}\}$, $B = \{\text{Tail on the third toss}\}$ and $C = \{\text{Exactly 2 tails in the 3 tosses}\}$.

1. Are events $A$ and $B$ independent?
2. Are events $B$ and $C$ independent?

# Random Variables

## Definition

A *random variable* is a (real-valued) **function**, $X$ that assigns a number to each sample point in $S$. Hence, $X : S \to \mathbb{R}$. Formally, the mapping need not be to $\mathbb{R}$ and both sets in the function $X$ need to be in a measurable space, i.e. associated with a $\sigma$-algebra.

- We denote the outcome of random variable $X$ as $x$, i.e. $X = x$
- A *discrete* random variable is associated with a finite or countably infinite sample space
- A *continuous* random variable is associated with an infinite (interval) sample space.

## Class Exercise

Consider the experiment of tossing a coin three times. How is the random variable, $X =$ no. of heads defined?

# Probability Density Function: Discrete Case

## Discrete Definition

A *probability density function, p.d.f*, (also called a *probability mass function, p.m.f.* in the discrete case) is a real valued function, $f$ that assigns a probability to the value of a discrete random variable $X$. Hence, $f : \mathbb{R} \to [0, 1]$. A more formal definition would involve a $\sigma$-algebra.

This means that for discrete random variables $f(x) = P(X = x)$ and $\sum_{i=1}^{\infty} f(x_i) = 1$.

## Class Exercise

Consider the experiment of tossing a coin three times and define the random variable as $X =$ no. of heads. Graph the p.d.f. of $X$.

# Probability Density Function: Continuous Case

## Continuous Definition

A *probability density function, p.d.f*, $f : \mathbb{R} \to A, A \subset \mathbb{R}$ denotes the ***likelihood*** that a continuous random variable, $X$ will take on a specific value. A more formal definition would involve a $\sigma$-algebra.

Hence, $f(x)$ does ***NOT*** give the value $P(X = x)$ since the total number of events are uncountable and hence, the probability that $X$ takes on any particular event is 0.

## Properties

The p.d.f. of a continuous random variable has the following properties:

- $f(x) \geq 0$
- $\int f(x)dx = 1$
- For any $a, b$ with $-\infty < a < b < \infty$, $P(a \leq X \leq b) = \int_a^b f(x)dx$.

# Cumulative Densities I

## Definition

A *cumulative density function, c.d.f.* is defined as $F(x) = P(X \leq x)$ for $-\infty < x < \infty$.

For the discrete case

$$F(x) = \sum_{t \leq x} f(t)$$

and for the continuous case

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

# Cumulative Densities II

## Class Exercise

Write out the c.d.f. for the previous coin tossing experiment.

## Class Exercise

Let $x$ be a continuous random variable with p.d.f $f(x) = 3e^{-3x}$ for $x \geq 0$.

1. Graph its p.d.f. and derive its c.d.f.
2. What is $P(0.5 \leq X \leq 1)$? What is $P(0.5 < X < 1)$?

# Joint Distributions

## Definitions

- For $X$ and $Y$, two discrete random variables, their *joint p.d.f* is such that $f(x, y) = P(X = x, Y = y)$
- The *joint c.d.f.* for the discrete case is $P(X \leq x, Y \leq y) = \sum_{s \leq x} \sum_{t \leq y} f(s, t)$
- For $X$ and $Y$, two continuous random variables, their *joint p.d.f.* is $P(X \in A, Y \in B) = \int_{x \in A} \int_{y \in B} f(x, y) dxdy$ for some sets $A$ and $B$.
- The *joint c.d.f.* for the continuous case is $F(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(s, t) dsdt$
- These definitions easily generalize to more than two variables.

# Marginal Distributions I

## Definition

The *marginal p.d.f* of $X_i$ is found by summing (integrating) out all the other $X_{-i}$ variables in the joint distribution for the discrete (continuous) case.

## Example

Suppose that there are $T$ people in the sample and that there are $a$ male smokers, $b$ female smokers, $c$ male nonsmokers and $d$ female nonsmokers. Let the random variable $X = 1$ if a randomly drawn person is female and 0 otherwise and let $Y = 1$ if a randomly drawn person is a smoker and 0 otherwise. Illustrate the joint p.d.f. of $X$ and $Y$ and calculate the marginal p.d.f. of $X$, $g(x)$.

# Marginal Distributions II

## Class Exercise

Given $f(x, y) = (2/3)(x + 2y)$ for $0 < x, y < 1$.

1. Calculate the marginal p.d.f. of $X$, $g(x)$.
2. Verify that $\int g(x)dx = 1$ and that $\int \int f(x, y)dxdy = 1$.

# Conditional Distributions

## Definition

The *conditional p.d.f* of two random variables $X$ and $Y$ is given as

$$f(x|y) = \frac{f(x,y)}{h(y)}$$

provided that $h(y) \neq 0$.

The above definition is valid for both discrete and continuous random variables and generalizes to more variables.

## Example

Compute $f(x|Y = 1)$ from the previous example.

## Class Exercise

Compute $P(X \leq 0.5|Y = 0.5)$ from the previous exercise.

# IID

## Definition

The previous definition of independent events can be extended to random variables. The random variables $X$ and $Y$ are *independent* if

$$f(x|y) = g(x).$$

Hence, $f(x, y) = g(x)h(y)$.

## Definition

We say that the random variables $X_i, i = 1, \ldots, n$ are *independent and identically distributed (IID)* if

$$f(x_1, \ldots, x_n) = f_1(x_1) \cdots f_n(x_n)$$

and

$$f_1(x_1) = f_2(x_2) = \cdots = f_n(x_n) = f(x).$$

# Change of Variable Technique I

- A function of a random variable is a random variable itself
- For instance $Y = g(X)$, where $X$ is a random variable
- To find the p.d.f of $Y$, $h(y)$ we make use of the *change of variable technique*
- This guarantees that the function integrates to one
- For the univariate case $h(y) = f[w(y)]|dw(y)/dy|$
- For the multivariate case the joint p.d.f $h(y_1, \ldots, y_n) = f[w_1, \ldots, w_n]|det(J)|$
- Certain requirements such as injectivity need to be met in order to use this method.

# Change of Variable Technique II

## Example

Let $f(x) = 2x$ for $0 < x < 1$ and let $Y = g(X) = 8X^3$.

1. Find $h(y)$ using a naïve method
2. Find $h(y)$ using the change of variable technique. What difference do you observe?

## Class Exercise

Let $f(x_1, x_2) = 1$ for $0 < x_1, x_2 < 1$ and let $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$. Find $h(y_1, y_2)$.

# Change of Variable Technique III

## Common Use

The change of variable technique is more commonly known for manipulations such as

$$Y = Z\sigma + \mu \sim N(\mu, \sigma^2),$$

where $Z \sim N(0, 1)$ is a standard normal distribution.

## Class Exercise

Verify the above result. Note that

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

# Expectation I

## Definition

The *expected* value of a random variable $X$, denoted $E[X]$ is the average or mean of $X$.

- For a discrete random variable $E[X] = \sum_x xf(x)$
- For a continuous random variable $E[X] = \int xf(x)dx$.

## Class Exercise

1. Find $E[X]$ for $X$ taking on the outcomes of tossing a fair coin.
2. Find $E[X]$ for $X$ taking on the outcomes of rolling a fair die.

# Expectation II

Expectation also holds for functions of random variables, for instance

$$E[g(X)] = \int g(x)f(x)dx$$

for the continuous case. Further, for some constants $a$ and $b$

$$E[aX + b] = aE[X] + b.$$

## Class Exercise

Show that the second property holds.

## Class Exercise

For $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$ and $g(X) = X^2$, find $E[g(X)]$.

# Jensen's Inequality I

## Definition

For the convex function $f$, *Jensen's inequality* states that

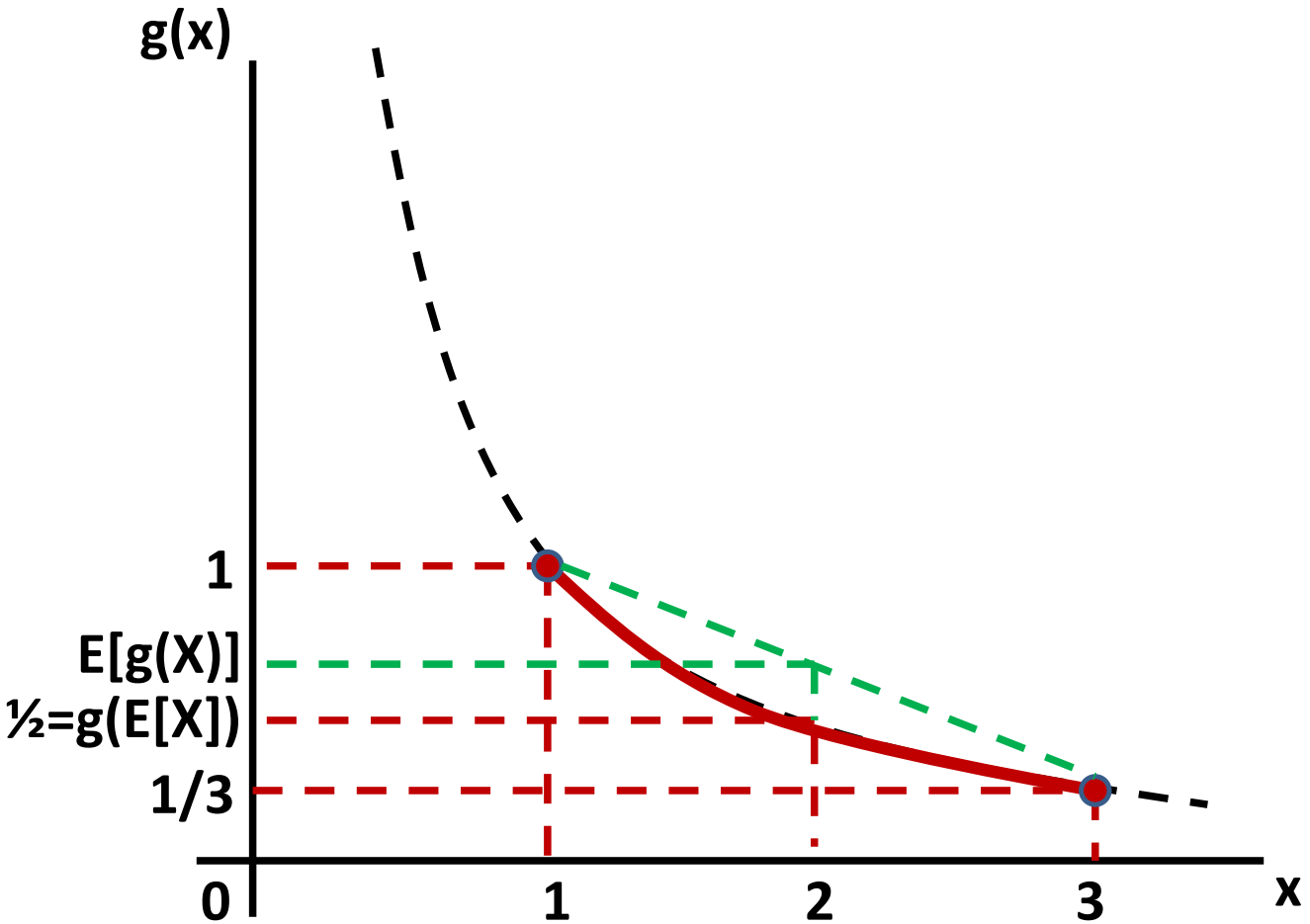
$$g(E[X]) \leq E[g(X)].$$

For a concave function, the inequality sign is reversed.

## Class Exercise

Let $X \sim U[1,3]$, show that Jensen's inequality holds for $g(X) = \frac{1}{X}$. **Hint**: the p.d.f. of a $U[a,b]$ random variable is $f(x) = \frac{1}{b-a}, x \in [a,b]$.

# Jensen's Inequality II



# Median and Mode

**Definition**

The *median* of a random variable is a value below which 50% of the probability mass falls.

**Definition**

The *mode* of a random variable is the peak (max) of the p.d.f.

**Note**

For a normally distributed random variable, the mean, median and the mode are the same.

# Conditional Expectation

## Definition

The *conditional expectation* of $X$ given $Y = y$ is

$$E[X|Y = y] = \int xf(x|y)dx.$$

# Law of Iterated Expectations

## Definition

For different realizations of $Y$, the conditional expectation will be a different number. Hence, we can view $E[X|Y]$ as a random variable. Suppose we take its expectation w.r.t. the distribution of $Y$:

$$E_Y[E[X|Y]] = E_X[X].$$

This result is known as the *law of iterated expectations*.

This result follows a similar reasoning used to derive the marginal distribution.

## Class Exercise

Derive the above result.

# Population Moments

## Definition

The *r*th *moment* of a random variable $X$ about the origin is $E[X^r]$. The *r*th moment about the mean of $X$ is $E[(X - E[X])^r]$.

- The mean of $X$ is therefore defined as the first moment about the origin
- The *variance* of $X$ is defined as the second moment about the mean
- The *covariance* of random variables $X$ and $Y$ is defined as the product of their first moments about the mean: $\text{Cov}(X, Y) = E[(X - E[X])]E[(Y - E[Y])]$
- The *correlation* between $X$ and $Y$ is given as

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} \leq |1|.$$

# Sample Moments

## Definition

Population moments are a theoretical concept. In practice we have a sample $(x_1, \ldots, x_T)$ from which we calculate the mean as

$$\bar{x} = \sum_{i=1}^{T} x_i / T$$

and the variance as

$$s_1^2 = \sum_{i=1}^{T} (x_i - \bar{x})^2 / T \quad \text{or} \quad s_2^2 = \sum_{i=1}^{T} (x_i - \bar{x})^2 / (T-1).$$

# Common Distributions

Some widely used distributions are

- Bernoulli and Binomial distributions (discrete)
- (Multivariate) Normal distribution
- $\chi^2$ distribution
- $t$ distribution
- $F$ distribution.

# Independence I

## Independence

Independence implies the absence of correlation,

$$\text{independence} \Rightarrow \text{no correlation.}$$

The converse is ***NOT*** true

$$\text{no correlation} \not\Rightarrow \text{independence.}$$

# Independence II

## Note

The multivariate normal distribution is a special case where no correlation implies independence.

## Class Exercise

Prove the above statement concerning the multivariate normal distribution. Note, an $(n \times 1)$ vector of normal random variables, $X \sim N(\mu, \mathbf{\Sigma})$ has the following density function

$$f(\mathbf{x}) = (2\pi)^{-n/2}|\mathbf{\Sigma}|^{-1/2}\exp\{(\mathbf{x}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)/2\}.$$

# Classical Estimation

- In classical statistical inference we tend to think in a *repeated sampling* sense.
- We assume that there is a *Data Generating Process (DGP)* that generates (infinitely) many data samples
- We then develop a rule (estimator) to estimate the sample parameters (eg. OLS)
- The properties of this estimator are analyzed, such as bias and precision
- This is of course within the context of *repeated sampling*
  - For a particular sample the estimator may miss the mark badly (depending on certain data characteristics)
  - Even though it is unbiased in a repeated sampling sense
- In practice we only have *one* sample.

# Estimator Properties: Small Samples

- In finite or so-called small samples we are concerned with estimator *bias*
- An unbiased estimator is such that on average (expectation) it will yield the true parameter value(s)
  - This means that if we had infinitely many samples, estimating the parameters of each sample and then taking the average of these estimates would yield the true parameter value(s)
- Another small sample property is estimator *efficiency*.
  - Out of two unbiased estimators, the one with the smaller variance is preferred.

# Estimator Properties: Large Samples (Asymptotics)

- In infinite or so-called large samples we are concerned with estimator *consistency*
  - A consistent estimator is such that with infinite sample size the estimator will *converge in probability* to the true parameter value, i.e.

$$\lim_{T \to \infty} P(|\widehat{\beta}_T - \beta| < \varepsilon) = 1$$

    no matter how small $\varepsilon > 0$ is
  - This means that as our sample size goes to infinity, we would estimate the true parameter values with our estimator
- Another large sample property is estimator *asymptotic efficiency*.
  - This is evaluated relative to the asymptotic information matrix.

# Convergence Results I

Some other popular convergence results used in econometrics are

- *Almost sure* convergence:

$$\lim_{T\to\infty} P(|\widehat{\beta}_T - \beta| = 0) = 1$$

- *Mean square* convergence:

$$\lim_{T\to\infty} E[(\widehat{\beta}_T - \beta)^2] < \varepsilon$$

- *Weak (Strong) Law of Large Numbers*: under fairly general conditions the sample mean converges in probability (almost surely) to the population mean
- *Central Limit Theorem*: under fairly general conditions the distribution of a random variable converges to the normal distribution.

# Convergence Results II

## Class Exercise

Show that the sample mean, $\bar{x_T} = \sum_{i=1}^{T} x_i / T$ converges in probability to the population mean, $\mu$. Assume that $x_i \sim$ iid$(\mu, \sigma^2)$ for $i = 1, 2, \ldots, T$. **Hint**: Show that it converges in mean square, this implies convergence in probability.

# End of Theme 5


DIW Berlin – Deutsches Institut
für Wirtschaftsforschung e.V.
Mohrenstraße 58, 10117 Berlin
www.diw.de