Week 01: Statistical Review

Mark M. Fredrickson (mfredric@umich.edu)

Computational Methods in Statistics and Data Science (Stats 406)

Probability and Random Variables

For an experiment that generates an **outcome**, the set of all possible results is called the **sample space** Ω.

Probability

For an experiment that generates an **outcome**, the set of all possible results is called the **sample space** Ω.

We say that **event** $A \subset \Omega$ has **probability** p if after **infinite repeated sampling** we observe that A occurs $100 \times p\%$ of the time.

Probability

For an experiment that generates an **outcome**, the set of all possible results is called the **sample space** Ω.

We say that **event** $A \subset \Omega$ has **probability** p if after **infinite repeated sampling** we observe that A occurs $100 \times p\%$ of the time.

Our notation is:

$$P(A) = p$$

Probability

For an experiment that generates an **outcome**, the set of all possible results is called the **sample space** Ω.

We say that **event** $A \subset \Omega$ has **probability** p if after **infinite repeated sampling** we observe that A occurs $100 \times p\%$ of the time.

Our notation is:

$$P(A) = p$$

Some example events:

- A coin coming up heads.
- Picking a blue ball and then a red ball from an box.
- The value of a stock exceeding \$100.
- Going bust or making over \$100 at the roulette table after 1 play.

Multiple Events, Independence

We often consider **multiple events**: $A \subset \Omega$ and $B \subset \Omega$.

Multiple Events, Independence

We often consider **multiple events**: $A \subset \Omega$ and $B \subset \Omega$.

We notate the **joint probability that both events A and B occur** as

$$P(A \text{ and } B) \equiv P(A \cap B) \equiv P(A, B)$$

Multiple Events, Independence

We often consider **multiple events**: $A \subset \Omega$ and $B \subset \Omega$.

We notate the **joint probability that both events A and B occur** as

$$P(A \text{ and } B) \equiv P(A \cap B) \equiv P(A, B)$$

Two events are **disjoint** if

$$P(A, B) = 0$$

Multiple Events, Independence

We often consider **multiple events**: $A \subset \Omega$ and $B \subset \Omega$.

We notate the **joint probability that both events A and B occur** as

$$P(A \text{ and } B) \equiv P(A \cap B) \equiv P(A, B)$$

Two events are **disjoint** if

$$P(A, B) = 0$$

Two events are **independent** if (and only if):

$$P(A, B) = P(A)P(B)$$

Conditional Probability

If we know B has occurred, the **conditional probability of A** is

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

Conditional Probability

If we know B has occurred, the **conditional probability of A** is

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} \Rightarrow P(A \text{ and } B) = P(A \mid B)P(B)$$

If we know B has occurred, the **conditional probability of A** is

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} \Rightarrow P(A \text{ and } B) = P(A \mid B)P(B)$$

Independence also implies

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

Random Variables

When outcomes can take values **on the real line** $\mathbb{R}$, we call them **random variables**.

Random Variables

When outcomes can take values **on the real line** $\mathbb{R}$, we call them **random variables**.

Often limit random variables to a subset of $\mathbb{R}$, which we call the support or domain: $\mathcal{D}$.

$$P(X \in \mathcal{D}) = 1 \iff P(X \notin \mathcal{D}) = 0$$

Random Variables

When outcomes can take values **on the real line** $\mathbb{R}$, we call them **random variables**.

Often limit random variables to a subset of $\mathbb{R}$, which we call the support or domain: $\mathcal{D}$.

$$P(X \in \mathcal{D}) = 1 \iff P(X \notin \mathcal{D}) = 0$$

If X can take any real value in $\mathcal{D}$, we say it is **continuous**.

Random Variables

When outcomes can take values **on the real line** $\mathbb{R}$, we call them **random variables**.

Often limit random variables to a subset of $\mathbb{R}$, which we call the support or domain: $\mathcal{D}$.

$$P(X \in \mathcal{D}) = 1 \iff P(X \notin \mathcal{D}) = 0$$

If X can take any real value in $\mathcal{D}$, we say it is **continuous**.

If X can only take a **countable** number of values (e.g., integers), it is **discrete**.

Random Variables

When outcomes can take values **on the real line** $\mathbb{R}$, we call them **random variables**.

Often limit random variables to a subset of $\mathbb{R}$, which we call the support or domain: $\mathcal{D}$.

$$P(X \in \mathcal{D}) = 1 \iff P(X \notin \mathcal{D}) = 0$$

If X can take any real value in $\mathcal{D}$, we say it is **continuous**.

If X can only take a **countable** number of values (e.g., integers), it is **discrete**.

Notation: uppercase X is the random variable, lower case x is a fixed value.

Distribution Functions

The **(cumulative) distribution function** (CDF) of a random variable is:

$$F(x) = P(X \leq x), \quad x \in \mathbb{R}$$

Distribution Functions

The **(cumulative) distribution function** (CDF) of a random variable is:

$$F(x) = P(X \leq x), \quad x \in \mathbb{R}$$

Useful properties:

- If $\mathcal{D} = [a, b]$, $x < a \Rightarrow F(x) = 0$ and $x \geq b \Rightarrow F(x) = 1$

The **(cumulative) distribution function** (CDF) of a random variable is:

$$F(x) = P(X \leq x), \quad x \in \mathbb{R}$$

Useful properties:

- If $\mathcal{D} = [a, b]$, $x < a \Rightarrow F(x) = 0$ and $x \geq b \Rightarrow F(x) = 1$
- More generally, $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

Distribution Functions

The **(cumulative) distribution function** (CDF) of a random variable is:

$$F(x) = P(X \leq x), \quad x \in \mathbb{R}$$

Useful properties:

- If $\mathcal{D} = [a, b]$, $x < a \Rightarrow F(x) = 0$ and $x \geq b \Rightarrow F(x) = 1$
- More generally, $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.
- F is **non-decreasing**: $F(x_1) \leq F(x_2)$ for $x_1 < x_2$.

Distribution Functions

The **(cumulative) distribution function** (CDF) of a random variable is:

$$F(x) = P(X \leq x), \quad x \in \mathbb{R}$$

Useful properties:

- If $\mathcal{D} = [a, b]$, $x < a \Rightarrow F(x) = 0$ and $x \geq b \Rightarrow F(x) = 1$
- More generally, $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.
- F is **non-decreasing**: $F(x_1) \leq F(x_2)$ for $x_1 < x_2$.
- F is **right continuous**: $\lim_{\epsilon \to 0^+} F(x + \epsilon) = F(x)$.

Continuous Distributions

If X is **continuous**, then F is a **continuous function**. If F is also **differentiable**, the **probability density function** (PDF) is:

$$f = \frac{d}{dx} F$$

Continuous Distributions

If X is **continuous**, then F is a **continuous function**. If F is also **differentiable**, the **probability density function** (PDF) is:

$$f = \frac{d}{dx} F$$

A consequence that if we have a region $\mathcal{R} \subset \mathcal{D}$,

$$P(X \in \mathcal{R}) = \int_{\mathcal{R}} f(x)\, dx$$

Continuous Distributions

If X is **continuous**, then F is a **continuous function**. If F is also **differentiable**, the **probability density function** (PDF) is:

$$f = \frac{d}{dx} F$$

A consequence that if we have a region $\mathcal{R} \subset \mathcal{D}$,

$$P(X \in \mathcal{R}) = \int_{\mathcal{R}} f(x)\, dx$$

For example:

$$P(X \leq t) = \int_{-\infty}^{t} f(x)\, dx = F(t)$$

Continuous Distributions

If X is **continuous**, then F is a **continuous function**. If F is also **differentiable**, the **probability density function** (PDF) is:

$$f = \frac{d}{dx}F$$

A consequence that if we have a region $\mathcal{R} \subset \mathcal{D}$,

$$P(X \in \mathcal{R}) = \int_{\mathcal{R}} f(x)\,dx$$

For example:

$$P(X \leq t) = \int_{-\infty}^{t} f(x)\,dx = F(t)$$

Note: we often suppress the fact that $f(x) = 0$ for $x \notin \mathcal{D}$

Example: $f(x) = 2(1-x)$

For a random variable X defined on support $0 \leq x \leq 1$, suppose the density is:

$$f(x) = 2(1-x)$$

Example: $f(x) = 2(1 - x)$

For a random variable X defined on support $0 \leq x \leq 1$, suppose the density is:

$$f(x) = 2(1 - x)$$

$$F(t) = \int_0^t 2 - 2x \, dx = t(2 - t)$$

Example: $f(x) = 2(1 - x)$

For a random variable X defined on support $0 \leq x \leq 1$, suppose the density is:

$$f(x) = 2(1 - x)$$

$$F(t) = \int_0^t 2 - 2x \, dx = t(2 - t)$$

$$P(X \leq 0.5) = \frac{3}{4}$$

Example: Normal Distribution

The **Normal distribution** ("Normal" is the name, not a descriptor) has PDF:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-(x-\mu)^2}{2\sigma^2} \right\}$$

with **parameters** are μ and σ^2.

Example: Normal Distribution

The **Normal distribution** ("Normal" is the name, not a descriptor) has PDF:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-(x-\mu)^2}{2\sigma^2} \right\}$$

with **parameters** are μ and σ^2.

We use the following notation to indicate that X is a Normal variable:

$$X \sim N(\mu, \sigma^2)$$

Example: Normal Distribution

The **Normal distribution** ("Normal" is the name, not a descriptor) has PDF:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-(x-\mu)^2}{2\sigma^2} \right\}$$

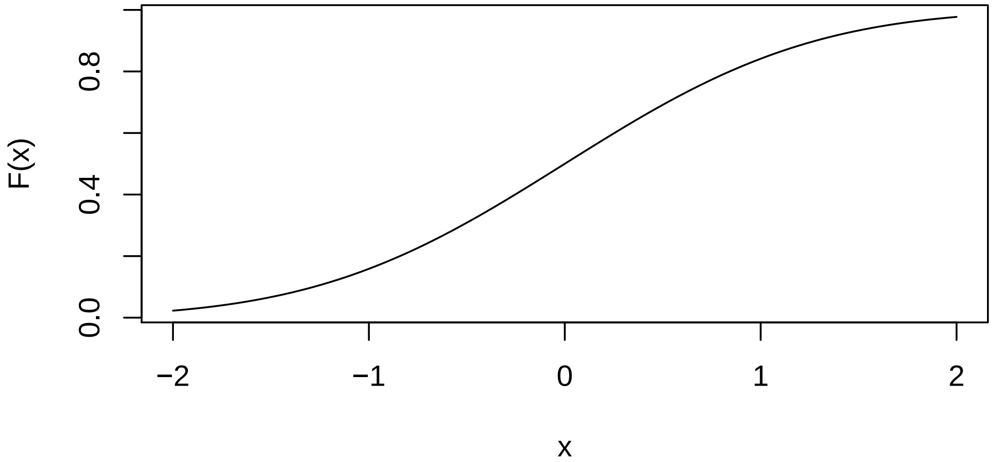
with **parameters** are μ and σ^2.

We use the following notation to indicate that X is a Normal variable:

$$X \sim N(\mu, \sigma^2)$$

There is no closed form for F, so we need to use look up tables that have been pre-computed using numerical procedures.

$N(0, 1)$



Discrete Distributions

If X is discrete (i.e., takes only values that can be mapped to the integers), then F is a **step function**.

If X is discrete (i.e., takes only values that can be mapped to the integers), then F is a **step function**.

Define the **probability mass function** (PMF) for X, as

$$f(x) = P(X = x)$$

If X is discrete (i.e., takes only values that can be mapped to the integers), then F is a **step function**.

Define the **probability mass function** (PMF) for X, as

$$f(x) = P(X = x)$$

As with the continuous case, we can build the CDF, from the PMF:

$$F(x) = \sum_{i=-\infty}^{x} P(X = i)$$

Bernoulli and Binomial Distributions

If X has **Bernoulli** distribution, it can take one value with probability θ and another value with probability $1-\theta$. Usually:

$$P(X=1)=\theta, \quad P(X=0)=1-\theta$$

Bernoulli and Binomial Distributions

If X has **Bernoulli** distribution, it can take one value with probability θ and another value with probability $1 - \theta$. Usually:

$$P(X = 1) = \theta, \quad P(X = 0) = 1 - \theta$$

We write:

$$X \sim \text{Bernoulli}(\theta)$$

Bernoulli and Binomial Distributions

If X has **Bernoulli** distribution, it can take one value with probability θ and another value with probability $1-\theta$. Usually:

$$P(X=1)=\theta, \quad P(X=0)=1-\theta$$

We write:

$$X \sim \text{Bernoulli}(\theta)$$

The **sum of n independent** Bernoulli variables has **Binomial** distribution:

$$X_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta), Y=\sum_{i=1}^{n} X_i \Rightarrow Y_i \sim \text{Binomial}(n,\theta)$$

Bernoulli and Binomial Distributions

If X has **Bernoulli** distribution, it can take one value with probability θ and another value with probability $1-\theta$. Usually:

$$P(X=1)=\theta, \quad P(X=0)=1-\theta$$

We write:

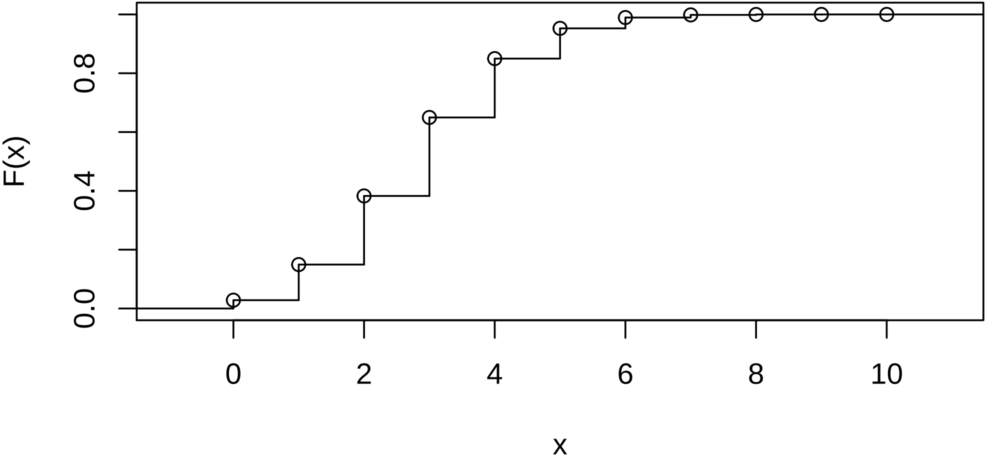
$$X \sim \text{Bernoulli}(\theta)$$

The **sum of n independent** Bernoulli variables has **Binomial** distribution:

$$X_i \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\theta),\, Y=\sum_{i=1}^{n} X_i \Rightarrow Y_i \sim \text{Binomial}(n,\theta)$$

$$P(Y=y)=\binom{n}{y}\theta^{y}(1-\theta)^{n-y}$$

Bernoulli(10, 0.3)



Joint Distributions

Random variables X and Y have a **joint CDF** described by:

$$F(x, y) = P(X \le x \text{ and } Y \le y)$$

Joint Distributions

Random variables X and Y have a **joint CDF** described by:

$$F(x, y) = P(X \leq x \text{ and } Y \leq y)$$

We write the **joint density or mass function** as $f(x, y)$.

Random variables X and Y have a **joint CDF** described by:

$$F(x, y) = P(X \leq x \text{ and } Y \leq y)$$

We write the **joint density or mass function** as $f(x, y)$.

Generally, the same properties hold for joint distributions as for univariate distributions. E.g.,

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$$

Example: Constrained support

Here is a density for RVs X and Y:

$$f(x,y) = \begin{cases} cx^2y & x^2 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- Compute c to make this a valid distribution
- Compute $P(X \geq Y)$

Finding c

By the **laws of total probability**,

$$1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \, dx \, dy$$

Finding c

By the **laws of total probability**,

$$1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy \quad = \int \int_{x^2 \le y \le 1} cx^2 y\, dx\, dy$$

Finding c

By the **laws of total probability**,

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy \quad = \int\int_{x^2 \le y \le 1} cx^2 y\,dx\,dy$$

$$= \int_{-1}^{1}\int_{x^2}^{1} cx^2 y\,dy\,dx$$

Finding c

By the **laws of total probability**,

$$
\begin{aligned}
1 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy &&= \int\int_{x^2\le y\le 1} cx^2y\,dx\,dy \\
&= \int_{-1}^{1}\int_{x^2}^{1} cx^2y\,dy\,dx &&= \int_{-1}^{1} cx^2\frac{1-x^4}{2}\,dx
\end{aligned}
$$

Finding c

By the **laws of total probability**,

$$
\begin{aligned}
1 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy &&= \int\int_{x^2\le y\le 1} cx^2y\,dx\,dy \\
&= \int_{-1}^{1}\int_{x^2}^{1} cx^2y\,dy\,dx &&= \int_{-1}^{1} cx^2\frac{1-x^4}{2}\,dx \\
&= c\left(\frac{x^3}{6}-\frac{x^7}{14}\Big|_{-1}^{1}\right)
\end{aligned}
$$

Finding c

By the **laws of total probability**,

$$\begin{aligned} 1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy \quad &= \int\int_{x^2\le y\le 1} cx^2 y\,dx\,dy \\ = \int_{-1}^{1}\int_{x^2}^{1} cx^2 y\,dy\,dx \quad &= \int_{-1}^{1} cx^2\frac{1-x^4}{2}\,dx \\ = c\left(\frac{x^3}{6}-\frac{x^7}{14}\Big|_{-1}^{1}\right) \quad &= c\frac{4}{21} \end{aligned}$$

So $c = 21/4$.

Example: $P(X \geq Y)$

Recall that the notation $P(X \geq Y)$ means, what is the probability of the **event that X is larger than Y**?

Example: $P(X \geq Y)$

Recall that the notation $P(X \geq Y)$ means, what is the probability of the **event that X is larger than Y**?

So we want to find all the probability contained in the region

$$\{(x, y) : x^2 \leq y \leq 1, x \geq y\}$$

Example: $P(X \geq Y)$

Recall that the notation $P(X \geq Y)$ means, what is the probability of the **event that X is larger than Y**?

So we want to find all the probability contained in the region

$$\{(x,y) : x^2 \leq y \leq 1, x \geq y\}$$

Notice: since $y \geq x^2$, it is also the case that $y \geq 0$. Therefore $x \geq 0$ for this set.

$$\int_0^1 \int_{x^2}^{x} \frac{21}{4} x^2 y \, dy \, dx = \frac{3}{20}$$

Marginal Distributions and Independence

We can **integrate out** one variable to get the **marginal distribution of the other**:

$$f(x) = \int_{-\infty}^{\infty} f(x, y)\, dy, \quad f(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$$

We can **integrate out** one variable to get the **marginal distribution of the other**:

$$f(x) = \int_{-\infty}^{\infty} f(x, y)\, dy, \quad f(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$$

E.g., immediately from previous example $f(x, y) = (21/4)x^2y$,

$$f(x) = \frac{21}{8}\left(x^2 - x^6\right), -1 \leq x \leq 1$$

Conditional Distributions

We will use the notation $X \mid Y = u$ to indicate a the random variable of X **conditioned on the event that** $Y = u$.

Conditional Distributions

We will use the notation $X \mid Y = u$ to indicate a the random variable of X **conditioned on the event that** $Y = u$.

Suppose that A is some event about X and B is the event $Y \in [u - \epsilon, u + \epsilon]$. Then,

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

Conditional Distributions

We will use the notation $X \mid Y = u$ to indicate a the random variable of X **conditioned on the event that** $Y = u$.

Suppose that A is some event about X and B is the event $Y \in [u - \epsilon, u + \epsilon]$. Then,

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{\int_A \int_{u-\epsilon}^{u+\epsilon} f_{xy}(x, y)\, dy\, dx}{\int_{u-\epsilon}^{u+\epsilon} f_y(y)\, dy}$$

Conditional Distributions

We will use the notation $X \mid Y = u$ to indicate a the random variable of X **conditioned on the event that** $Y = u$.

Suppose that A is some event about X and B is the event $Y \in [u - \epsilon, u + \epsilon]$. Then,

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{\int_A \int_{u-\epsilon}^{u+\epsilon} f_{xy}(x, y)\, dy\, dx}{\int_{u-\epsilon}^{u+\epsilon} f_y(y)\, dy}$$

Taking the limit as $\epsilon \to 0$, we get the **conditional density (or mass) function for** $X \mid Y = u$:

$$f(x \mid y = u) = \frac{f_{xy}(x, u)}{f_y(u)}$$

Example: Conditional Distribution

Suppose

$$f(x, y) = \theta^2 (xy)^{\theta-1}, \quad 0 \leq x \leq 1, 0 \leq y \leq 1, \theta \geq 0$$

Example: Conditional Distribution

Suppose

$$f(x,y) = \theta^2 (xy)^{\theta-1}, \quad 0 \le x \le 1, 0 \le y \le 1, \theta \ge 0$$

$$f(x) = \theta x^{\theta-1} \int_0^1 \theta y^{\theta-1}\, dy = \theta x^{\theta-1} \left(y^\theta \Big|_0^1 \right) = \theta x^{\theta-1}$$

Example: Conditional Distribution

Suppose

$$f(x, y) = \theta^2 (xy)^{\theta-1}, \quad 0 \leq x \leq 1, 0 \leq y \leq 1, \theta \geq 0$$

$$f(x) = \theta x^{\theta-1} \int_0^1 \theta y^{\theta-1}\, dy = \theta x^{\theta-1} \left(y^{\theta} \Big|_0^1 \right) = \theta x^{\theta-1}$$

$$f(y \mid x) = \frac{f(x, y)}{f(x)} = \theta y^{\theta-1}$$

Factoring Independent RVs

Suppose that for any y, the conditional distribution of X is the same as it's marginal distribution: $f(x \mid y) = f(x)$.

Factoring Independent RVs

Suppose that for any y, the conditional distribution of X is the same as it's marginal distribution: $f(x \mid y) = f(x)$.

Then it must be the case that

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

Factoring Independent RVs

Suppose that for any y, the conditional distribution of X is the same as it's marginal distribution: $f(x \mid y) = f(x)$.

Then it must be the case that

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

Factoring Independent RVs

Suppose that for any y, the conditional distribution of X is the same as it's marginal distribution: $f(x \mid y) = f(x)$.

Then it must be the case that

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

In other words, X and Y are **independent** by our earlier definition.

Factoring Independent RVs

Suppose that for any y, the conditional distribution of X is the same as it's marginal distribution: $f(x \mid y) = f(x)$.

Then it must be the case that

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

In other words, X and Y are **independent** by our earlier definition.

Factorizing the joint density (mass) function is both **necessary and sufficient** for independence.

Factoring Independent RVs

Suppose that for any y, the conditional distribution of X is the same as it's marginal distribution: $f(x \mid y) = f(x)$.

Then it must be the case that

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

In other words, X and Y are **independent** by our earlier definition.

Factorizing the joint density (mass) function is both **necessary and sufficient** for independence.

This result also applies to CDFs: $F_{xy}(a,b) = F_x(a)F_y(b)$

Expectation

Suppose we are going to compute $g(X)$ for a random variable X.

Expectation

Suppose we are going to compute $g(X)$ for a random variable X.

We often want to "average over" X to get a sense of a typical value for $g(X)$. We define the **expectation** of $g(X)$ as:

$$E(g(X)) = \sum_{i=-\infty}^{\infty} P(X = x)g(x) \qquad \text{(discrete)}$$

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)\,dx \qquad \text{(continuous)}$$

We call the **expectation of the identity function**, $E(X)$ the **mean of X**.

Expectation

Suppose we are going to compute $g(X)$ for a random variable X.

We often want to "average over" X to get a sense of a typical value for $g(X)$. We define the **expectation** of $g(X)$ as:

$$E(g(X)) = \sum_{i=-\infty}^{\infty} P(X = x)g(x) \qquad \text{(discrete)}$$

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)\,dx \qquad \text{(continuous)}$$

We call the **expectation of the identity function**, $E(X)$ the **mean of X**.

We call $E\left[(X - E(X))^2\right]$ the **variance**.

Example: Computing $E(Y)$ for Bernoulli(10, 0.3)

Recall that $f(y)$ is

$$P(Y = y) = \binom{10}{y} (0.3)^y (0.7)^{10-y}$$

and the support is the integers from zero to ten.

Example: Computing $E(Y)$ for Bernoulli(10, 0.3)

Recall that $f(y)$ is

$$P(Y = y) = \binom{10}{y}(0.3)^y(0.7)^{10-y}$$

and the support is the integers from zero to ten.

$$E(Y) = \sum_{i=0}^{10}\binom{10}{i}(0.3)^i(0.7)^{10-i} \times i$$

Example: Computing $E(Y)$ for Bernoulli(10, 0.3)

Recall that $f(y)$ is

$$P(Y = y) = \binom{10}{y}(0.3)^y(0.7)^{10-y}$$

and the support is the integers from zero to ten.

$$E(Y) = \sum_{i=0}^{10} \binom{10}{i}(0.3)^i(0.7)^{10-i} \times i$$

```
> terms <- map_dbl(0:10, function(i) {
+   choose(10, i) * 0.3^i * 0.7^(10 - i) * i
+ })
> sum(terms)

[1] 3
```

Example: Computing $E(\log(Y+1))$ for Bernoulli(10, 0.3)

The function $g(x) = \log(x+1)$ is defined for 0 to 10, so we can ask:

$$E(\log(Y+1)) = ?$$

Example: Computing $E(\log(Y+1))$ for Bernoulli(10, 0.3)

The function $g(x) = \log(x+1)$ is defined for 0 to 10, so we can ask:

$$E(\log(Y+1)) = ?$$

```
> terms <- map_dbl(0:10, function(i) {
+   choose(10, i) * 0.3^i * 0.7^(10 - i) * log(i + 1)
+ })
> sum(terms)

[1] 1.311
```

Example: Expectation for a Continuous RV

Suppose we have

$$f(x) = \theta x^{\theta-1}, \quad 0 \leq x \leq 1, \theta > 0$$

and we want to find **the variance of X**, $\mathrm{E}\left(X^2\right) - \mathrm{E}\left(X\right)^2$.

Example: Expectation for a Continuous RV

Suppose we have

$$f(x) = \theta x^{\theta-1}, \quad 0 \leq x \leq 1, \theta > 0$$

and we want to find **the variance of X**, $\mathrm{E}\left(X^2\right) - \mathrm{E}\left(X\right)^2$.

$$\mathrm{E}\left(X\right) = \int_0^1 x f(x)\, dx$$

Example: Expectation for a Continuous RV

Suppose we have

$$f(x) = \theta x^{\theta-1}, \quad 0 \leq x \leq 1, \theta > 0$$

and we want to find **the variance of X**, $\mathrm{E}\left(X^2\right) - \mathrm{E}\left(X\right)^2$.

$$\mathrm{E}\left(X\right) = \int_0^1 x f(x)\, dx = \int_0^1 \theta x^{\theta}\, dx = \frac{\theta}{\theta+1}$$

Example: Expectation for a Continuous RV

Suppose we have

$$f(x) = \theta x^{\theta-1}, \quad 0 \leq x \leq 1, \theta > 0$$

and we want to find **the variance of X**, $\mathrm{E}\left(X^2\right) - \mathrm{E}\left(X\right)^2$.

$$\mathrm{E}\left(X\right) = \int_0^1 xf(x)\,dx = \int_0^1 \theta x^{\theta}\,dx = \frac{\theta}{\theta+1}$$

$$\mathrm{E}\left(X^2\right) = \int_0^1 \theta x^{\theta+1}\,dx = \frac{\theta}{\theta+2}$$

Example: Expectation for a Continuous RV

Suppose we have

$$f(x) = \theta x^{\theta-1}, \quad 0 \leq x \leq 1, \theta > 0$$

and we want to find **the variance of X**, $\mathrm{E}\left(X^2\right) - \mathrm{E}\left(X\right)^2$.

$$\mathrm{E}\left(X\right) = \int_0^1 x f(x)\, dx = \int_0^1 \theta x^{\theta}\, dx = \frac{\theta}{\theta+1}$$

$$\mathrm{E}\left(X^2\right) = \int_0^1 \theta x^{\theta+1}\, dx = \frac{\theta}{\theta+2}$$

$$\mathrm{Var}\left(X\right) = \frac{\theta}{(\theta+2)} - \frac{\theta^2}{(\theta+1)^2}$$

Conditional Expectation

Recall that $X \mid Y = y$ is a random variable, so it we can consider **the conditional expectation of X given $Y = y$**:

$$E(X \mid Y = y) = \int_{-\infty}^{\infty} x\, f(x \mid y)\, dx$$

Conditional Expectation

Recall that $X \mid Y = y$ is a random variable, so it we can consider **the conditional expectation of X given $Y = y$**:

$$\mathrm{E}\left(X \mid Y = y\right) = \int_{-\infty}^{\infty} x\, f(x \mid y)\, dx$$

The result will be a function of y, i.e., $h(y) = \int_{-\infty}^{\infty} x\, f(x \mid y)\, dx$. This leads to the useful result of the **law of iterated expectation**:

$$\mathrm{E}\left(h(Y)\right) = \mathrm{E}\left(\mathrm{E}\left(X \mid Y\right)\right) = \mathrm{E}\left(X\right)$$

Properties of Expectations

Some useful facts (which also apply to $g(X)$, $h(Y)$, etc):

- $E(aX + b) = aE(X) + b$

Properties of Expectations

Some useful facts (which also apply to $g(X)$, $h(Y)$, etc):

- $E(aX + b) = aE(X) + b$
- $E(X + Y) = E(X) + E(Y)$

Properties of Expectations

Some useful facts (which also apply to $g(X)$, $h(Y)$, etc):

- $E(aX + b) = aE(X) + b$
- $E(X + Y) = E(X) + E(Y)$
- If X and Y are **independent**, then $E(XY) = E(X)E(Y)$ (**the converse is not true!**)

Properties of Expectations

Some useful facts (which also apply to $g(X)$, $h(Y)$, etc):

- $E(aX + b) = aE(X) + b$
- $E(X + Y) = E(X) + E(Y)$
- If X and Y are **independent**, then $E(XY) = E(X)E(Y)$ (**the converse is not true!**)
- $\text{Var}(aX + b) = a^2\text{Var}(X)$

Properties of Expectations

Some useful facts (which also apply to $g(X)$, $h(Y)$, etc):

- $E(aX + b) = aE(X) + b$
- $E(X + Y) = E(X) + E(Y)$
- If X and Y are **independent**, then $E(XY) = E(X)E(Y)$ (**the converse is not true!**)
- $\text{Var}(aX + b) = a^2\text{Var}(X)$
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$ where

$$\text{Cov}(X, Y) = E\left[(X - E(X))(Y - E(Y))\right] = E(XY) - E(X)E(Y)$$

Properties of Expectations

Some useful facts (which also apply to $g(X)$, $h(Y)$, etc):

- $E(aX + b) = aE(X) + b$
- $E(X + Y) = E(X) + E(Y)$
- If X and Y are **independent**, then $E(XY) = E(X)E(Y)$ (**the converse is not true!**)
- $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$ where

$$\mathrm{Cov}(X, Y) = E\left[(X - E(X))(Y - E(Y))\right] = E(XY) - E(X)E(Y)$$

- If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$

Analytic solution for Binomial mean

Recall we can think of $Y \sim \text{Binomial}(n, \theta)$:

$$Y = \sum_{i=1}^{n} X_i, \quad X_1, X_2, \dots X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$$

Analytic solution for Binomial mean

Recall we can think of $Y \sim \text{Binomial}(n, \theta)$:

$$Y = \sum_{i=1}^{n} X_i, \quad X_1, X_2, \ldots X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$$

Then

$$E(Y) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

Analytic solution for Binomial mean

Recall we can think of $Y \sim \text{Binomial}(n, \theta)$:

$$Y = \sum_{i=1}^{n} X_i, \quad X_1, X_2, \ldots X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$$

Then

$$E(Y) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

Since $E(X_1) = E(X_2) = \ldots = E(X_n)$, further:

$$E(Y) = nE(X_1)$$

Analytic solution for Binomial mean

Recall we can think of $Y \sim \text{Binomial}(n, \theta)$:

$$Y = \sum_{i=1}^{n} X_i, \quad X_1, X_2, \ldots X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$$

Then

$$E(Y) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

Since $E(X_1) = E(X_2) = \ldots = E(X_n)$, further:

$$E(Y) = nE(X_1)$$

Finally,

$$E(X_1) = \theta \times 1 + (1 - \theta) \times 0 = \theta \Rightarrow E(Y) = n\theta$$

- **Random variables** are random outcomes described by **real numbers**
- All RVs have **(cumulative) distribution function**: $F(x) = \Pr X \leq x$
- **Continuous** RV: (a) probability density functions $f(x)$, (b) **expectation** is $E(g(X)) = \int_{-\infty}^{\infty} g(x) f(x)\, dx$.
- **Discrete** RVs: (a) probability mass functions $p(x)$, (b) expectation is $E(g(X)) = \sum_{x \in \Omega} p(x) x$
- **Independence**: the joint distribution is the production of the marginal distributions.

Inference

What is inference?

In the previous examples we **posited a model** (i.e., assumed a distribution) for data.

What is inference?

In the previous examples we **posited a model** (i.e., assumed a distribution) for data.

Typically, we leave some parts of the model **unknown**. We call these unknown components **parameters**.

What is inference?

In the previous examples we **posited a model** (i.e., assumed a distribution) for data.

Typically, we leave some parts of the model **unknown**. We call these unknown components **parameters**.

After we **encounter data**, we want to make **reasonable guesses** (i.e., estimates) or **validate possible values** (i.e., test hypotheses) for the parameters.

What is inference?

In the previous examples we **posited a model** (i.e., assumed a distribution) for data.

Typically, we leave some parts of the model **unknown**. We call these unknown components **parameters**.

After we **encounter data**, we want to make **reasonable guesses** (i.e., estimates) or **validate possible values** (i.e., test hypotheses) for the parameters.

We call these process **inference**. We want to tools that **behave well** when performing inference (i.e., operating characteristics).

Statistics

A **statistic** is a function of **random variables**.

A **statistic** is a function of **random variables**.

Example: the **sample mean**:

$$T(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Statistics

A **statistic** is a function of **random variables**.

Example: the **sample mean**:

$$T(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$$

Statistics

A **statistic** is a function of **random variables**.

Example: the **sample mean**:

$$T(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$$

We use statistics to

- **Reduce the size** of our data. If we can do so without losing information we call them **sufficient**.
- **Estimate** parameters of a population from a sample.
- **Test** statistical hypotheses about how our data were generated.

Statistics

A **statistic** is a function of **random variables**.

Example: the **sample mean**:

$$T(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$$

We use statistics to

- **Reduce the size** of our data. If we can do so without losing information we call them **sufficient**.
- **Estimate** parameters of a population from a sample.
- **Test** statistical hypotheses about how our data were generated.

Important: statistics are **random variables** too!

Estimation

If $X_1, X_2, \ldots X_n$ are from the **same distribution**, we say that they are **identical**. Often, we also assume **independence** (IID):

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F(\theta)$$

where θ can be a **vector of many parameters**.

Estimation

If $X_1, X_2, \ldots X_n$ are from the **same distribution**, we say that they are **identical**. Often, we also assume **independence** (IID):

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F(\theta)$$

where θ can be a **vector of many parameters**.

An easy way to get IID is to sample from a **large, well defined** population uniformly at random (**simple random sample**).

If $X_1, X_2, \dots X_n$ are from the **same distribution**, we say that they are **identical**. Often, we also assume **independence** (IID):

$$X_1, X_2, \dots, X_n \overset{\text{iid}}{\sim} F(\theta)$$

where θ can be a **vector of many parameters**.

An easy way to get IID is to sample from a **large, well defined** population uniformly at random (**simple random sample**).

We often wish to estimate θ for the population using an **estimator** (a statistic):

$$\hat{\theta}(X_1, X_2, \dots, X_n) = \text{ do something with the } X_i \text{ values}$$

Sampling Distributions

Since $\hat{\theta}$ is a **random variable** it has a distribution. We call it the **sampling distribution**.

Sampling Distributions

Since $\hat{\theta}$ is a **random variable** it has a distribution. We call it the **sampling distribution**.

Example: if $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, then $\bar{X}_n \sim N(\mu, \sigma^2/n)$

Since $\hat{\theta}$ is a **random variable** it has a distribution. We call it the **sampling distribution**.

Example: if $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, then $\bar{X}_n \sim N(\mu, \sigma^2/n)$

Generally, we want to know some **properties of an estimator**:

- **Bias**: $E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$

Since $\hat{\theta}$ is a **random variable** it has a distribution. We call it the **sampling distribution**.

Example: if $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, then $\bar{X}_n \sim N(\mu, \sigma^2/n)$

Generally, we want to know some **properties of an estimator**:

- **Bias**: $E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$
- **Variance**: $\text{Var}(\hat{\theta})$

Since $\hat{\theta}$ is a **random variable** it has a distribution. We call it the **sampling distribution**.

Example: if $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, then $\bar{X}_n \sim N(\mu, \sigma^2/n)$

Generally, we want to know some **properties of an estimator**:

- **Bias**: $E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$
- **Variance**: $\text{Var}(\hat{\theta})$
- **Mean Squared Error (MSE)**:

$$E\left[(\hat{\theta} - \theta)^2\right] = \text{Bias}^2 + \text{Var}(\hat{\theta})$$

Example: Sample Mean of $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$

We will use without proof the fact that $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2) \Rightarrow \bar{X}_n \sim N(\mu, \sigma^2/n)$.

Example: Sample Mean of $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$

We will use without proof the fact that $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2) \Rightarrow \bar{X}_n \sim N(\mu, \sigma^2/n)$.

With respect to μ, what is the bias and variance of the estimator

- $\hat{\mu} = X_1$: $\mathrm{E}\left(X_1\right) - \mu = \mu - \mu = 0$, $\mathrm{Var}\left(X_1\right) = \sigma^2$

Example: Sample Mean of $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$

We will use without proof the fact that $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2) \Rightarrow \bar{X}_n \sim N(\mu, \sigma^2/n)$.

With respect to μ, what is the bias and variance of the estimator

- $\hat{\mu} = X_1$: $\mathrm{E}\left(X_1\right) - \mu = \mu - \mu = 0$, $\mathrm{Var}\left(X_1\right) = \sigma^2$
- $\hat{\mu} = \bar{X}_n$: $\mathrm{E}\left(\bar{X}_n\right) = \mu - \mu = 0$, $\mathrm{Var}\left(\bar{X}_n\right) = \sigma^2/n$.

Example: Sample Mean of $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$

We will use without proof the fact that $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2) \Rightarrow \bar{X}_n \sim N(\mu, \sigma^2/n)$.

With respect to μ, what is the bias and variance of the estimator

- $\hat{\mu} = X_1$: $\mathrm{E}\left(X_1\right) - \mu = \mu - \mu = 0$, $\mathrm{Var}\left(X_1\right) = \sigma^2$
- $\hat{\mu} = \bar{X}_n$: $\mathrm{E}\left(\bar{X}_n\right) = \mu - \mu = 0$, $\mathrm{Var}\left(\bar{X}_n\right) = \sigma^2/n$.

Conclusion, the **MSE of a single observation is n times larger than the MSE of $\bar{X}$**.

Method of Moments Estimation

We call expectations of the form $E(X^r)$ the **moments of X**. E.g., the mean is the first moment.

Method of Moments Estimation

We call expectations of the form $E(X^r)$ the **moments of X**. E.g., the mean is the first moment.

Sometimes we can write $\theta = g(\mathrm{E}\left(X^1\right), \mathrm{E}\left(X^2\right), \ldots, \mathrm{E}\left(X^r\right))$ for some r and some g.

Method of Moments Estimation

We call expectations of the form $E(X^r)$ the **moments of X**. E.g., the mean is the first moment.

Sometimes we can write $\theta = g(\mathrm{E}\left(X^1\right), \mathrm{E}\left(X^2\right), \ldots, \mathrm{E}\left(X^r\right))$ for some r and some g.

We can estimate $E(X^r)$ using the **sample moments**:

$$m_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r$$

and solve $\hat{\theta} = g(m_1, m_2, \ldots, m_r)$ for $\hat{\theta}$ (perhaps using a system of equations).

Method of Moments Estimation

We call expectations of the form $E(X^r)$ the **moments of X**. E.g., the mean is the first moment.

Sometimes we can write $\theta = g(\mathrm{E}\left(X^1\right), \mathrm{E}\left(X^2\right), \ldots, \mathrm{E}\left(X^r\right))$ for some r and some g.

We can estimate $E(X^r)$ using the **sample moments**:

$$m_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r$$

and solve $\hat{\theta} = g(m_1, m_2, \ldots, m_r)$ for $\hat{\theta}$ (perhaps using a system of equations).

Example: $f(x) = \theta x^{\theta-1}$

For $f(x) = \theta x^{\theta-1}$, we previously computed

$$\mathrm{E}(X) = \frac{\theta}{\theta + 1}$$

Example: $f(x) = \theta x^{\theta - 1}$

For $f(x) = \theta x^{\theta-1}$, we previously computed

$$\mathrm{E}(X) = \frac{\theta}{\theta + 1}$$

Substituting $\bar{X}$ for $\mathrm{E}(X)$, and solving for θ we get:

$$\bar{X} = \frac{\hat{\theta}}{\hat{\theta} + 1} \Rightarrow \hat{\theta} = \frac{\bar{X}}{1 - \bar{X}}$$

Likelihood Functions

A sample has **joint density/mass function** as:

$$f(x_1, x_2, \ldots, x_n; \theta)$$

Likelihood Functions

A sample has **joint density/mass function** as:

$$f(x_1, x_2, \ldots, x_n; \theta)$$

Instead of thinking about the x_i **values as arguments**, we can think about θ as the argument to get the **likelihood function**:

$$L(\theta; x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n; \theta)$$

Likelihood Functions

A sample has **joint density/mass function** as:

$$f(x_1, x_2, \ldots, x_n; \theta)$$

Instead of thinking about the x_i **values as arguments**, we can think about θ as the argument to get the **likelihood function**:

$$L(\theta; x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n; \theta)$$

Note: When X_i **are IID**,

$$L(\theta; x_1, x_2, \ldots, x_n) = \prod f(x_i; \theta)$$

Maximum Likelihood Estimation

With a likelihood function $L(\theta)$, we can ask, "What θ would seem to make my data most plausible?"

Maximum Likelihood Estimation

With a likelihood function $L(\theta)$, we can ask, "What θ would seem to make my data most plausible?"

This implies that we should find the **maximum likelihood estimator**:

$$\hat{\theta} = \text{argmax}_{\theta \in \Theta} L(\theta; x_1, x_2, \ldots, x_n)$$

Maximum Likelihood Estimation

With a likelihood function $L(\theta)$, we can ask, "What θ would seem to make my data most plausible?"

This implies that we should find the **maximum likelihood estimator**:

$$\hat{\theta} = \text{argmax}_{\theta \in \Theta} L(\theta; x_1, x_2, \ldots, x_n)$$

We can often achieve this goal by maximizing a **monotonic transformation**, such as the log function.

Maximum Likelihood Estimation

With a likelihood function $L(\theta)$, we can ask, "What θ would seem to make my data most plausible?"

This implies that we should find the **maximum likelihood estimator**:

$$\hat{\theta} = \mathrm{argmax}_{\theta \in \Theta} L(\theta; x_1, x_2, \ldots, x_n)$$

We can often achieve this goal by maximizing a **monotonic transformation**, such as the log function.

MLEs have many nice properties including **invariance** and **low variance**.

Logarithmic transformations

Recall the definition of the logarithm for base b,

$$\log_b(x) = a : b^a = x$$

Logarithmic transformations

Recall the definition of the logarithm for base b,

$$\log_b(x) = a : b^a = x$$

In this course, we'll always take log be the "natural logarithm", $\log_e$ (though usually the base doesn't matter due to cancellation) .

Logarithmic transformations

Recall the definition of the logarithm for base b,

$$\log_b(x) = a : b^a = x$$

In this course, we'll always take log be the "natural logarithm", $\log_e$ (though usually the base doesn't matter due to cancellation) .

Some useful things to remember:

- $\log(e^x) = x$
- $\exp(x + y) = \exp(x)\exp(y)$, so $\log(xy) = \log(x) + \log(y)$,
- $\log(x^y) = y \log(x)$, with previous we get $\log(x/y) = \log(x) - \log(y)$
- $$\frac{d}{dx}\log(x) = \frac{1}{x}$$

Example: Normal mean, $\sigma^2 = 1$

The likelihood for μ in $N(\mu, 1)$ is

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x_i - \mu)^2\right\}$$

Example: Normal mean, $\sigma^2 = 1$

The likelihood for μ in $N(\mu, 1)$ is

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x_i - \mu)^2\right\}$$

Taking the log likelihood yields:

$$\log(L(\mu)) = l(\mu) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}(x_i - \mu)^2$$

Example: Normal mean, $\sigma^2 = 1$

The likelihood for μ in $N(\mu, 1)$ is

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x_i - \mu)^2\right\}$$

Taking the log likelihood yields:

$$\log(L(\mu)) = l(\mu) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}(x_i - \mu)^2$$

Our standard calculus strategy is to take the derivative and set to zero:

$$0 = \sum_{i=1}^{n} -(x_i - \mu) \Rightarrow n\mu = \sum_{i=1}^{n} x_i \Rightarrow \hat{\mu} = \bar{X}$$

Example: $f(x) = \theta x^{\theta-1}$

Suppose we have n independent samples, each with distribution function $f(x) = \theta x^{\theta-1}$.

Example: $f(x) = \theta x^{\theta-1}$

Suppose we have n independent samples, each with distribution function $f(x) = \theta x^{\theta-1}$.

$$L(\theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1} = \theta^n \left(\prod_{i=1}^{n} x_i \right)^{\theta-1}$$

Example: $f(x) = \theta x^{\theta-1}$

Suppose we have n independent samples, each with distribution function $f(x) = \theta x^{\theta-1}$.

$$L(\theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1} = \theta^n \left(\prod_{i=1}^{n} x_i \right)^{\theta-1}$$

Taking the log,

$$l(\theta) = n \log(\theta) + (\theta - 1) \sum_{i=1}^{n} \log(x_i)$$

Example: $f(x) = \theta x^{\theta-1}$

Suppose we have n independent samples, each with distribution function $f(x) = \theta x^{\theta-1}$.

$$L(\theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1} = \theta^n \left(\prod_{i=1}^{n} x_i \right)^{\theta-1}$$

Taking the log,

$$l(\theta) = n\log(\theta) + (\theta - 1)\sum_{i=1}^{n} \log(x_i)$$

And the derivative **with respect to θ** and setting to zero:

$$\frac{d}{d\theta} l(\theta) = \frac{n}{\theta} + \sum_{i=1}^{n} \log(x_i)$$

Example: $f(x) = \theta x^{\theta-1}$

Suppose we have n independent samples, each with distribution function $f(x) = \theta x^{\theta-1}$.

$$L(\theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1} = \theta^n \left(\prod_{i=1}^{n} x_i \right)^{\theta-1}$$

Taking the log,

$$l(\theta) = n \log(\theta) + (\theta - 1) \sum_{i=1}^{n} \log(x_i)$$

And the derivative **with respect to θ** and setting to zero:

$$\frac{d}{d\theta} l(\theta) = \frac{n}{\theta} + \sum_{i=1}^{n} \log(x_i) \Rightarrow \hat{\theta} = \frac{-n}{\sum_{i=1}^{n} \log(x_i)}$$

NB: notice that for $x \leq 1$, $\log(x) \leq \log(1) = 0$, so $\hat{\theta} \geq 0$.

Hypothesis Tests

Another place we use statistics is for **hypothesis tests**.

Hypothesis Tests

Another place we use statistics is for **hypothesis tests**.

A hypothesis test requires stating a **null hypothesis** H_0 and an **alternative hypothesis** H_1. Some examples:

$$H_0 : X \overset{\text{iid}}{\sim} F_0 \qquad \text{vs.} \quad H_1 : X \overset{\text{iid}}{\sim} F_1$$

Hypothesis Tests

Another place we use statistics is for **hypothesis tests**.

A hypothesis test requires stating a **null hypothesis** H_0 and an **alternative hypothesis** H_1. Some examples:

$$H_0 : X \overset{\text{iid}}{\sim} F_0 \qquad \text{vs.} \quad H_1 : X \overset{\text{iid}}{\sim} F_1$$

$$H_0 : E(X) \leq \mu_0 \qquad \text{vs.} \quad H_1 : E(X) > \mu_0$$

Another place we use statistics is for **hypothesis tests**.

A hypothesis test requires stating a **null hypothesis** H_0 and an **alternative hypothesis** H_1. Some examples:

$$
\begin{aligned}
H_0 :& X \overset{\text{iid}}{\sim} F_0 && \text{vs. } H_1 : X \overset{\text{iid}}{\sim} F_1 \\
H_0 :& E(X) \leq \mu_0 && \text{vs. } H_1 : E(X) > \mu_0 \\
H_0 :& F(x, y) = F_x(x) F_x(y) && \text{v.s. } H_1 : F(x, y) \neq F_x(x) F_y(y)
\end{aligned}
$$

Hypothesis Tests

Another place we use statistics is for **hypothesis tests**.

A hypothesis test requires stating a **null hypothesis** H_0 and an **alternative hypothesis** H_1. Some examples:

$$\begin{aligned} H_0 :& X \overset{\text{iid}}{\sim} F_0 && \text{vs. } H_1 : X \overset{\text{iid}}{\sim} F_1 \\ H_0 :& E(X) \leq \mu_0 && \text{vs. } H_1 : E(X) > \mu_0 \\ H_0 :& F(x, y) = F_x(x)F_x(y) && \text{v.s. } H_1 : F(x, y) \neq F_x(x)F_y(y) \end{aligned}$$

Goal: Either **accept** the null hypothesis or **reject** the null hypothesis in favor of the alternative.

Type I and Type II Error

If we **reject a true null hypothesis**, we have committed a **Type I error**.

Type I and Type II Error

If we **reject a true null hypothesis**, we have committed a **Type I error**.

If we **accept a false null hypothesis**, we have committed a **Type II error**.

Type I and Type II Error

If we **reject a true null hypothesis**, we have committed a **Type I error**.

If we **accept a false null hypothesis**, we have committed a **Type II error**.

The probability of making a Type I error is the **size** of the test:

$$P(\text{Reject } H_0 \mid H_0)$$

Type I and Type II Error

If we **reject a true null hypothesis**, we have committed a **Type I error**.

If we **accept a false null hypothesis**, we have committed a **Type II error**.

The probability of making a Type I error is the **size** of the test:

$$P(\text{Reject } H_0 \mid H_0)$$

We define the probability of **not making a Type II error** when H_1 is true as the **power** of the test:

$$P(\text{Reject } H_0 \mid H_1)$$

Type I and Type II Error

If we **reject a true null hypothesis**, we have committed a **Type I error**.

If we **accept a false null hypothesis**, we have committed a **Type II error**.

The probability of making a Type I error is the **size** of the test:

$$P(\text{Reject } H_0 \mid H_0)$$

We define the probability of **not making a Type II error** when H_1 is true as the **power** of the test:

$$P(\text{Reject } H_0 \mid H_1)$$

Useful framework: pick a **maximum Type I error** α and then pick a test that has **good power**.

Test Statistics

In the previous slide we defined size and power using the **probability of rejecting** the null hypothesis (when it was true or false, respectively).

Test Statistics

In the previous slide we defined size and power using the **probability of rejecting** the null hypothesis (when it was true or false, respectively).

This probability comes from the **test statistic** we use to make our decision:

$$T(X_1, \ldots, X_n)$$

Test Statistics

In the previous slide we defined size and power using the **probability of rejecting** the null hypothesis (when it was true or false, respectively).

This probability comes from the **test statistic** we use to make our decision:

$$T(X_1, \ldots, X_n) = T$$

Test Statistics

In the previous slide we defined size and power using the **probability of rejecting** the null hypothesis (when it was true or false, respectively).

This probability comes from the **test statistic** we use to make our decision:

$$T(X_1, \ldots, X_n) = T$$

and a **rejection region** $\mathcal{R}$ such that

$$T \in \mathcal{R} \iff \text{reject the null hypothesis}$$

Test Statistics

In the previous slide we defined size and power using the **probability of rejecting** the null hypothesis (when it was true or false, respectively).

This probability comes from the **test statistic** we use to make our decision:

$$T(X_1, \ldots, X_n) = T$$

and a **rejection region** $\mathcal{R}$ such that

$$T \in \mathcal{R} \iff \text{reject the null hypothesis}$$

Usually, we pick $\mathcal{R}$ so that we maintain our α-level:

$$P(T \in \mathcal{R} \,|\, H_0) \leq \alpha \quad \text{(size less than level)}$$

and generates high power:

$$P(T \in \mathcal{R} \,|\, H_1) \text{ is large}$$

Example: Testing $\mu_0 = 0$ vs. $\mu_0 = 1$

Suppose we assume that

$$X_i \overset{\text{iid}}{\sim} N(\mu, 1) \quad (\mu \text{ unknown})$$

and want to test:

$$H_0 : \mu = 0 \quad \text{vs. } H_1 : \mu = 1$$

Example: Testing $\mu_0 = 0$ vs. $\mu_0 = 1$

Suppose we assume that

$$X_i \overset{\text{iid}}{\sim} N(\mu, 1) \quad (\mu \text{ unknown})$$

and want to test:

$$H_0 : \mu = 0 \quad \text{vs. } H_1 : \mu = 1$$

We already know that $\bar{X}_n$ is the MLE for μ, perhaps that would be a good statistic:

- When H_0 is true,

$$\bar{X}_n \sim N\left(0, \frac{1}{n}\right)$$

Example: Testing $\mu_0 = 0$ vs. $\mu_0 = 1$

Suppose we assume that

$$X_i \overset{\text{iid}}{\sim} N(\mu, 1) \quad (\mu \text{ unknown})$$

and want to test:

$$H_0 : \mu = 0 \quad \text{vs. } H_1 : \mu = 1$$

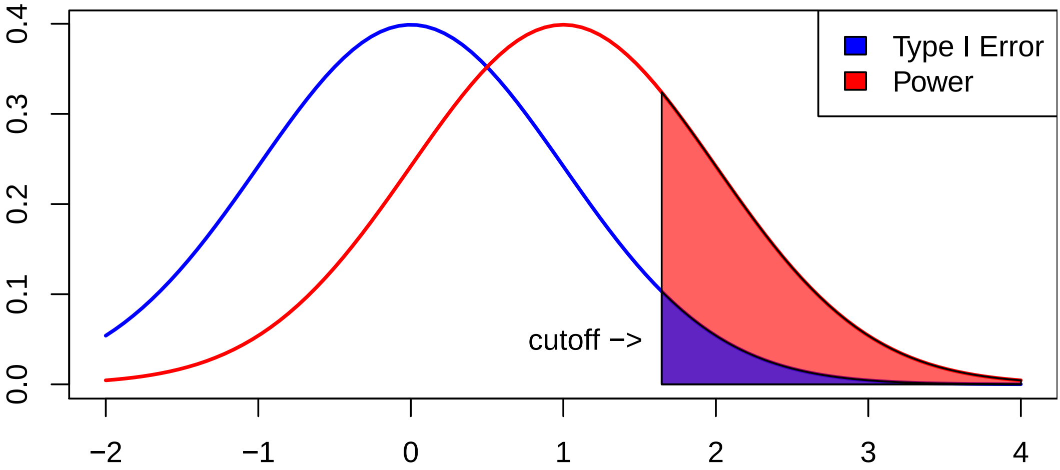
We already know that $\bar{X}_n$ is the MLE for μ, perhaps that would be a good statistic:

- When H_0 is true,

$$\bar{X}_n \sim N\left(0, \frac{1}{n}\right)$$

- When H_1 is true,

$$\bar{X}_n \sim N\left(1, \frac{1}{n}\right)$$



Type I Error
Power
cutoff ->
0.0
0.1
0.2
0.3
0.4
−2
−1
0
1
2
3
4

Computing rejection region and power ($n = 2$)

Computing the rejection region when $H_0 : \mu = 0$:

```
> n <- 2
> (cutoff <- qnorm(0.95, mean = 0, sd = 1/n))

[1] 0.8224
```

Computing rejection region and power ($n = 2$)

Computing the rejection region when $H_0 : \mu = 0$:

```
> n <- 2
> (cutoff <- qnorm(0.95, mean = 0, sd = 1/n))
```

```
[1] 0.8224
```

Computing the power of the test when $H_1 : \mu = 1$:

```
> 1 - pnorm(cutoff, mean = 1, sd = 1/n)
```

```
[1] 0.6388
```

Summary: Inference

- Write down quantities of interest as **population parameters**.
- Use **sample statistics** make decisions about parameters.
- **Estimation**: make reasonable guess, **sampling distribution** defines uncertainty
- **Hypothesis tests**: see if data conform to hypothesis, **null** and **alternative** distributions define uncertainty.
- **Method of moments** and **maximum likelihood** will be our two main estimation techniques.