# 36-401, Chapter 1: Review of Random Variables

Zach Branson, Fall 2025

## Motivation

**Regression** involves learning the relationship (if any) between **outcomes** $Y$ and **covariates** $X$. Given $n$ observations, our data will look like $(X_1, Y_1), \ldots, (X_n, Y_n)$, where each $X_i$ and $Y_i$ is organized by columns in a dataset.

**Example 0.1.** The US Bureau of Economic Analysis (BEA) releases data on the economic output of metropolitan areas. Below we consider data from 2006, where we assess the relationship between **per-capita gross metropolitan product** (GMP) (the outcome) and **population size** (the covariate).

We can load the data and look at the first few rows:

```
bea <- read.csv("data/bea-2006.csv")
head(bea)
```

```
                                Y      X
##                         MSA pcgmp    pop finance prof.tech     ict ma
## 1               Abilene, TX 24490 158700 0.09750        NA 0.01621
## 2                 Akron, OH 32890 699300 0.12940   0.05440      NA
## 3                Albany, GA 24270 163000 0.08217        NA 0.00708
```

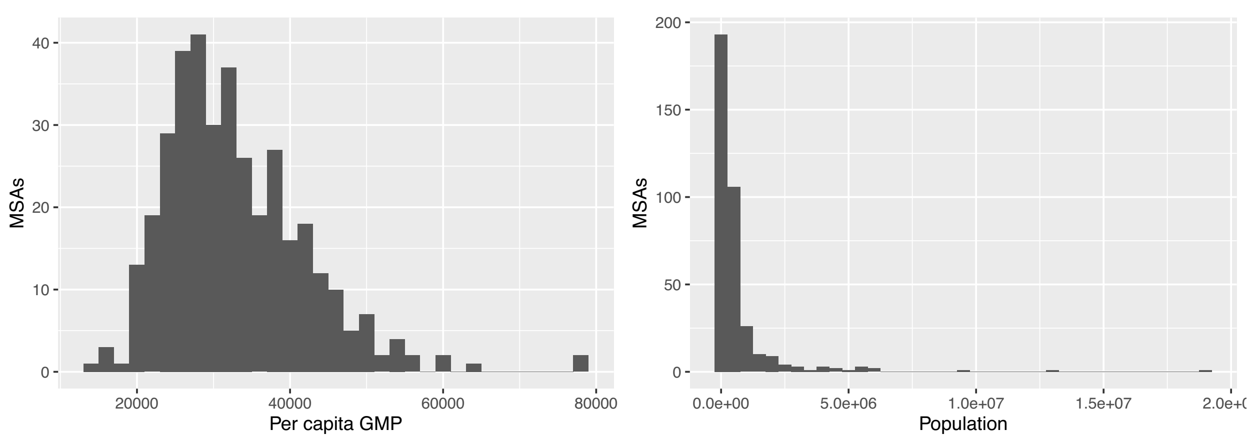
```
## 4 Albany-Schenectady-Troy, NY 36840 850300 0.15780   0.09399 0.04511
## 5             Albuquerque, NM 37660 816000 0.15990   0.09978 0.20500
## 6              Alexandria, LA 25490 152200 0.09152   0.03790 0.01134
```

We'll visualization `pcgmp`, `pop`, and their relationship.

First, we can visualize the variables' *marginal distributions*.

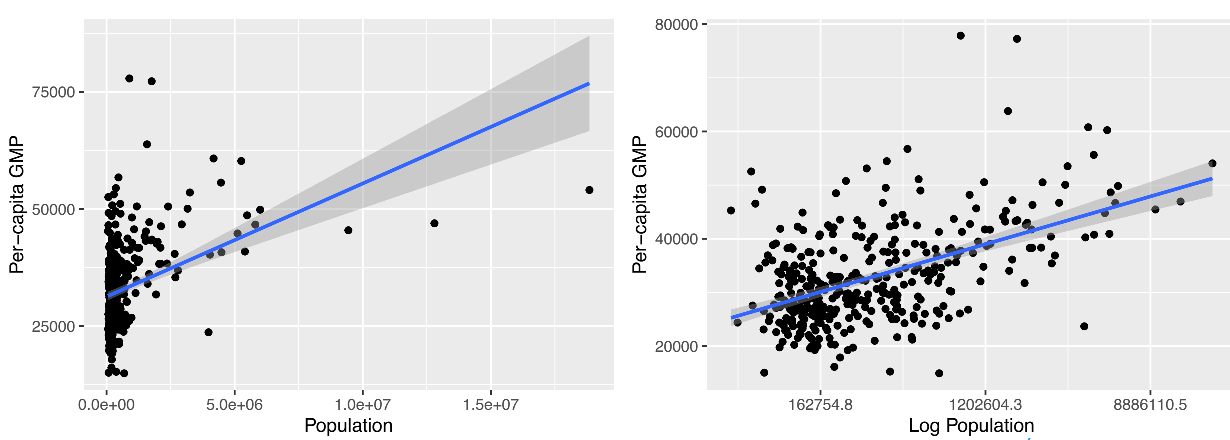
```r
library(ggplot2)
library(gridExtra)
pcgmpHist <- ggplot(bea, aes(x = pcgmp)) +
  geom_histogram(binwidth = 2000) +
  labs(x = "Per capita GMP", y = "MSAs")
popHist <- ggplot(bea, aes(x = pop)) +
  geom_histogram(binwidth = 500000) +
  labs(x = "Population", y = "MSAs")

grid.arrange(pcgmpHist, popHist, ncol = 2)
```



Now, we can visualize the variables' *joint distribution* and their *linear relationship* with a scatterplot and linear regression line. Below we use pop as a covariate (left) or log(pop) as a covariate (right).

```
#population as covariate
popScatter <- ggplot(bea, aes(x = pop, y = pcgmp)) +
  geom_point() +
  geom_smooth(method = "lm") + # linear model plotted on top
  labs(x = "Population", y = "Per-capita GMP")
#log(population) as covariate
logPopScatter <- ggplot(bea, aes(x = pop, y = pcgmp)) +
  geom_point() +
  scale_x_continuous(trans = "log") +
  geom_smooth(method = "lm") +
  labs(x = "Log Population", y = "Per-capita GMP")
grid.arrange(popScatter, logPopScatter, ncol = 2)
```




doesn't seem linear

Y vs X

seems more linear

Y vs logX

There are several key takeaways from the above example.

- $X$ & $Y$ are random variables w/ distributions

  Randomness comes from sampling $(X_1, Y_1), \ldots (X_n, Y_n)$

- Can consider marginal distribution, joint distr, and conditional distr (eg. $Y \mid X$)

- Linear regression lines plots $\hat{E}[Y \mid X] = \hat{\beta}_0 + \hat{\beta}_1 X$

- Estimators $\hat{\beta}_0$, $\hat{\beta}_1$ are functions of $(X, Y)$ and thus are random variables.

- We'll consider expectation, variance, covariance, distr. of rand. vars. to conduct inference.

Unknown Population Quantities

$E[X], E[Y], E[Y \mid X]$

data-generating process

Sample

$(X_1, Y_1), \ldots, (X_n, Y_n)$

Modeling analysis

Estimation

$\hat{E}[X]$, $\hat{E}[Y]$, $\hat{E}[Y \mid X]$

stat. Inference

# Distributions of Random Variables

A **discrete random variable** $X$ is characterized by its **probability mass function (pmf)**, denoted $f(\cdot)$, where

$$f(k) = \mathbb{P}(X = k) \quad \text{for all } k.$$

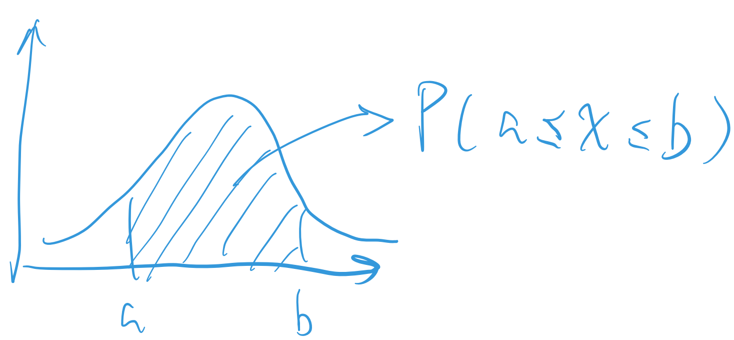
Here, $k$ corresponds to different potential values of $X$.

**Exercise 0.1.** List some properties that the pmf function must possess.

- $\sum_k f(k) = 1$
- $0 \leq \underbrace{f(k)}_{P(X=k)} \leq 1$ for all $k$

A **continuous random variable** $X$ is characterized by its **probability density function (pdf)**, denoted $f(x)$, such that

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\,dx \quad \text{for all } a \leq b.$$



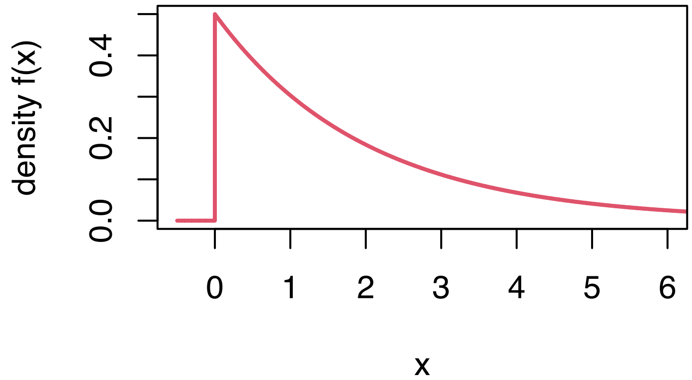



Figure 0.1: The Exponential($\lambda$) pdf when $\lambda = 0.5$.

**Example 0.2.** A random variable $X$ is said to have the **Exponential**($\lambda$) distribution if

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

**Exercise 0.2.** Suppose $X \sim \text{Exp}(\lambda)$.
What is $\mathbb{P}(X \leq a)$? What is $\mathbb{P}(3 \leq X \leq 5)$? What is $\mathbb{P}(X = 2)$?

$P(X \leq a) = \int_0^a f(x)\, dx$, where $f(x) = \lambda e^{-\lambda x}$

$P(3 \leq X \leq 5) = \int_3^5 f(x)\, dx$

$P(X = 2) = 0$, $\left( \int_2^2 f(x)\, dx = 0 \right)$

# Expected Values

The **expected value** of a random variable $X$ is defined as

$$\mathbb{E}(X) = \sum_k k f(k) \qquad \text{when } X \text{ is discrete, and}$$

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

Heuristically, $\mathbb{E}[X]$ is the "average" value a random variable $X$ takes. Often, expectations are denoted with $\mu$.

The **variance** of $X$ is defined as

$$\text{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

The variance measures the spread of a distribution, often denoted with $\sigma^2$. The square root of the variance is the **standard deviation**.

The **covariance** between $X$ and $Y$ is defined as

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

Covariance measures the strength of linear relationship between $X$ and $Y$.

A related quantity is the **correlation** between $X$ and $Y$, defined as

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

$\sigma_x = \sqrt{var(x)}$
$\sigma_Y = \sqrt{var(Y)}$

The correlation is bounded between $-1$ and $1$.

There are several important properties of expectations, variances, and covariances that we'll use throughout this class.

- Linearity of $E[\cdot]$: $E[aX+bY+c] = aE[X]+bE[Y]+c$
  for scalers $a,b,c$
- Law of Unconscious Statistician (LOTUS)
  $$E[h(x)] = \int_{-\infty}^{\infty} h(x)\, f(x)\, dx$$
- $Var(aX+b) = a^2 Var(X)$
- $Var(aX+bY) = a^2 Var(X) + b^2 Var(Y) + 2ab\, Cov(X,Y)$
- $Cov(aX+b, Y) = a\, Cov(X,Y)$
- $Cov(aX+bY, cU+dV) = ac\, Cov(X,U) + ad\, Cov(X,V) + bc\, Cov(Y,U) + bd\, Cov(Y,V)$

**Exercise 0.3.** What is $\text{Cov}(X, X)$?

$$\begin{aligned} Cov(X, X) &= E[X \cdot X] - E[X] \cdot E[X] \\ &= E[X^2] - (E[X])^2 \\ &= Var(X) \end{aligned}$$

The above expectations, variances, covariances, and correlations are **population-level** quantities that we'll estimate with sample analogs from the data.

- Sample Avg. : $\hat{E}[X] = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Sample Var. : $\widehat{Var}(X) = S_X^2 = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(X_i - \bar{X})^2$
- Sample Cov. : $\widehat{Cov}(X, Y) = S_{XY} = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$
- Sample Correlation : $\widehat{Corr}(X, Y) = r_{XY} = \frac{S_{X,Y}}{S_X, S_Y}$

The above sample analogs are all **statistics**: They are functions of the data. They each have a **sampling distribution**: i.e., their distribution when we repeatedly obtain many samples of size $n$.

$\underbrace{\text{statistic}}_{\text{random var.}}$ : $T = g((X_1, Y_1), \ldots, (X_n, Y_n))$

sampling distr : distr. of $T$ across samples

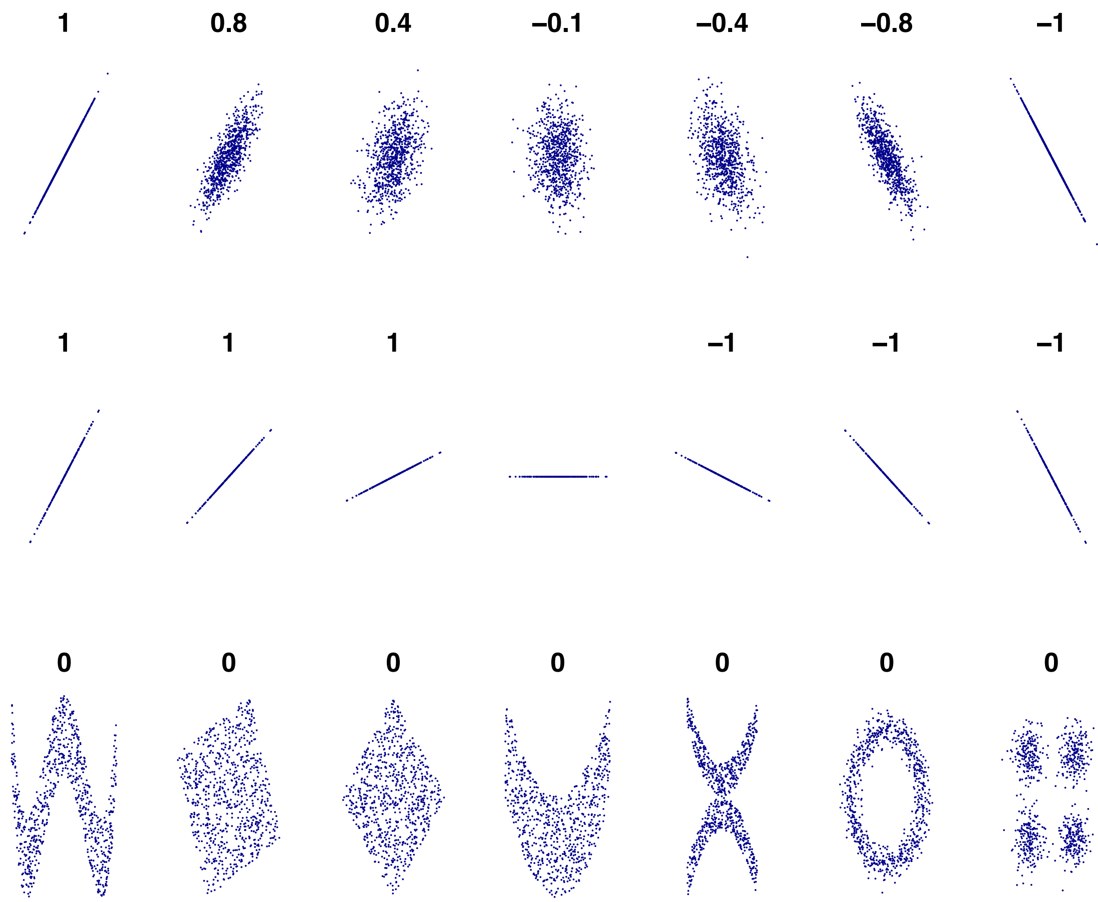



Figure 0.2: Examples of scatter plots, and corresponding correlations $r$. From *Wikipedia*.

Figure 0.2 shows some example scatterplots and corresponding correlations. It is useful for building intuition about correlation values.

# Conditional Expectation

Conditional expectations let us ask: What is the mean of $Y$ among observations where $X = x$?

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

$$\mathbb{E}(Y|X=x) = \begin{cases} \sum_y y f(y|x) & \text{discrete case} \\ \int y f(y|x) dy & \text{continuous case.} \end{cases}$$

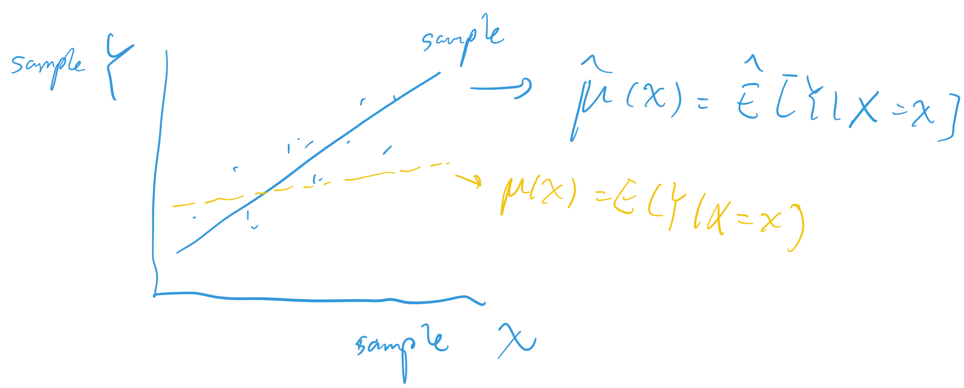
This is the same definition of expectation, but we replaced the marginal $f_Y(y)$ with the conditional $f_{Y|X}(y|x)$.

Similarly, if $r(x,y)$ is a function of $x$ and $y$, then

$$\mathbb{E}(r(X,Y)|X=x) = \begin{cases} \sum_y r(x,y) f(y|x) & \text{discrete case} \\ \int r(x,y) f(y|x) dy & \text{continuous case.} \end{cases}$$

**Warning!** Whereas $\mathbb{E}(Y)$ is a number, $\mathbb{E}(Y|X=x)$ is a *function* of $x$. In fact, $\mathbb{E}(Y|X)$ is a random variable whose value is $\mathbb{E}(Y|X=x)$ when $X = x$.

$E[Y \mid \underbrace{X}_{\text{rand. var.}} = x] = \mu(x)$ (func. of $x$)



**Exercise 0.4.** Suppose we draw $X \sim \text{Unif}(0, 1)$. After we observe $X = x$, we draw $Y \mid X = x \sim \text{Unif}(x, 1)$. What is $\mathbb{E}(Y \mid X)$?

In general, $Z \sim \text{Unif}(a, b)$, then: $f(z) = \begin{cases} \frac{1}{b-a} & \text{for } a < z < b \\ 0 & \text{otherwise} \end{cases}$

Thus: $f(y|x) = \begin{cases} \frac{1}{1-x} & \text{for } x < y < 1 \\ 0 & \text{otherwise} \end{cases}$

$$E[Y | X = x] = \int y \cdot f(y|x)\, dy$$

$$= \frac{1}{1-x} \int_x^1 y\, dy$$

$$= \left(\frac{1}{1-x}\right) \cdot \left(\frac{1}{2} y^2 \Big|_x^1\right)$$

$$= \left(\frac{1}{1-x}\right)\left(\frac{1}{2} - \frac{1}{2}x^2\right)$$

$$= \frac{1}{2(1-x)} (1-x) \cdot (1+x)$$

$$= \frac{1+x}{2}$$

The **conditional variance** of $Y$ given $X = x$ is

$$\text{Var}(Y|X = x) = \int (y - \mu(x))^2 f(y|x) dy$$

where $\mu(x) = \mathbb{E}(Y|X = x)$.

Again, the conditional variance is a function of $x$ and a random variable: It is the variance of $Y$ when (by chance) $X = x$.

Two very important properties of conditional expectations and variances:

**Exercise 0.5.** Return to the above example. What is $\mathbb{E}(Y)$?

# Large-Sample Theorems

In this class, we'll consider the *asymptotic* (i.e., large-sample) behavior of estimators in terms of their *bias* and *variance*.

The below two foundational results establish the asymptotic behavior of the sample mean $\overline{X}$ as an estimator for $\mathbb{E}[X]$.

**The Law of Large Numbers.** Assume $(X_1, \ldots, X_n)$ are independent and identically distributed (iid), where $\mathbb{E}[X_i] < \infty$. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mathbb{E}(X).$$

The $\xrightarrow{p}$ means "convergence in probability." Informally, this means that the bias and variance of $\overline{X}$ go to zero as $n \to \infty$. Formally, this means

$$\lim_{n\to\infty} P(|\overline{X} - \mathbb{E}[X]| > \epsilon) = 0, \text{ for any } \epsilon > 0$$

Thus, the sample mean is a *consistent estimator* for the population mean, which is reassuring.

**Central Limit Theorem.** Assume $(X_1, \ldots, X_n)$ are iid, where $\mathbb{E}[X_i] < \infty$ and $\mathrm{Var}(X_i) < \infty$. Then, as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mathbb{E}[X], \frac{\mathrm{Var}(X)}{n}\right)$$

Mathematically, it is nicer to have the limit that we're converging to not change with $n$. Thus, the CLT is often stated as

$$\sqrt{n}\left(\frac{\overline{X} - \mathbb{E}(X)}{\sqrt{\mathrm{Var}(X)}}\right) \xrightarrow{d} \mathcal{N}(0,1).$$

The $\xrightarrow{d}$ means "convergence in distribution." The CLT tells us that not only is $\overline{X}$ an unbiased estimator, but also it has an asymptotically Normal distribution with a defined variance. If we can estimate this variance, then the distribution provides a way to compute confidence intervals.

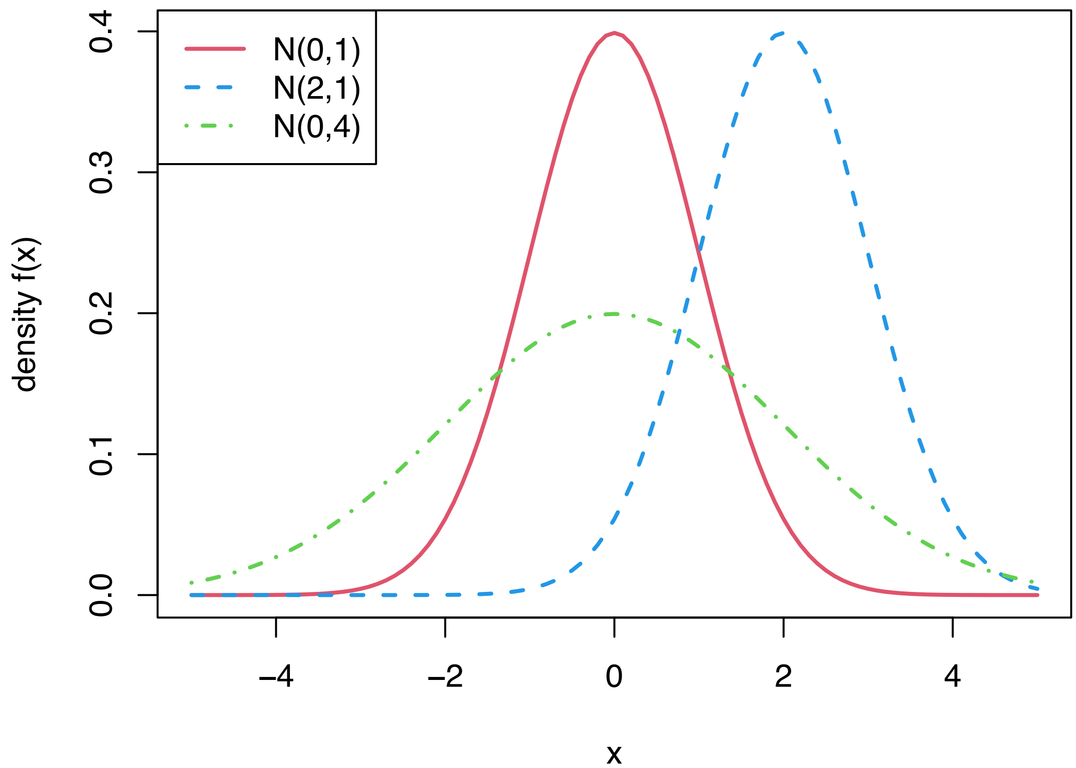



Figure 0.3: Three Normal densities.

## The Normal Distribution

The **Normal distribution** has the classic bell-shaped density that we have learned to love, shown in Figure 0.3:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty$$

Writing $X \sim N(\mu, \sigma^2)$, implies that $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$.

The case where $\mu = 0$ and $\sigma^2 = 1$ is called the **standard Normal**.

**Exercise 0.6.** Suppose that $X_1, X_2, \ldots, X_n$ are each Normally distributed, i.e., $X_i \sim N(\mu_i, \sigma_i^2)$. Under what condition(s) is the linear combination

$$Y = \sum_{i=1}^{n} a_i X_i$$

(where at least one $a_i \neq 0$) also Normal?

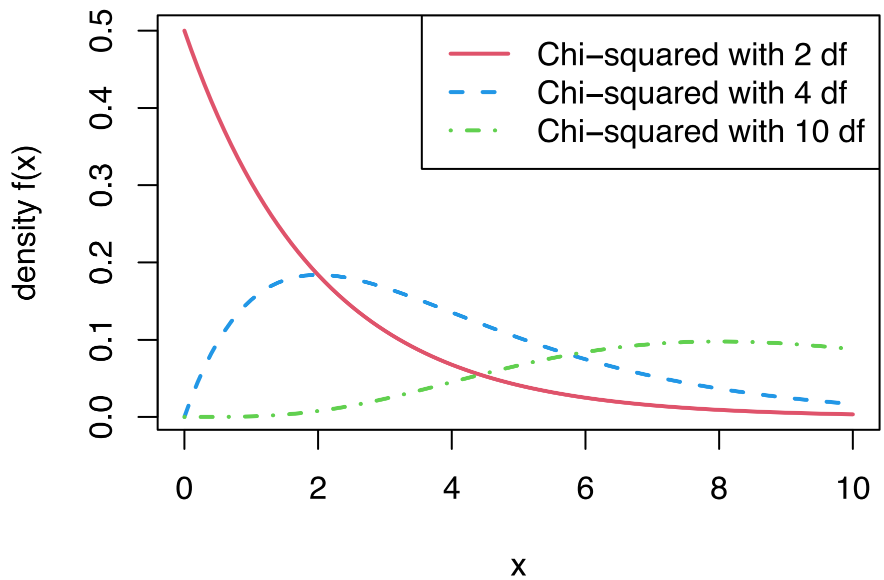



Figure 0.4: Three chi-squared densities.

# Other Important Distributions

## Chi-Squared Distribution

If $Z_1, Z_2, \ldots, Z_n \overset{iid}{\sim} N(0,1)$, then

$$X = \sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$$

i.e., the above sum follows a **chi-squared distribution with $n$ degrees of freedom**.

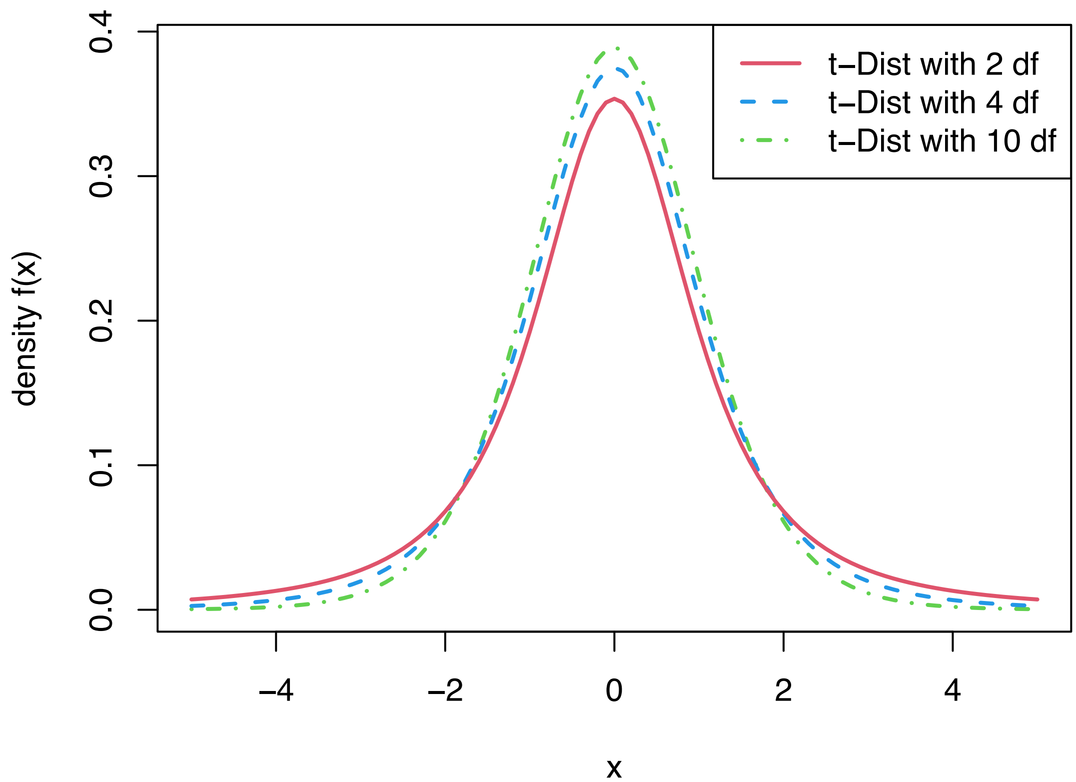



Figure 0.5: Three t-distribution densities.

## The t-Distribution

If $Y \sim N(0,1)$ and $U \sim \chi^2_n$ independent of $Y$, then

$$X = Y \bigg/ \sqrt{\frac{U}{n}} \sim t_n$$

i.e., the above quantity follows a **t-distribution with** $n$ **degrees of freedom**. This distribution has a Normal-like shape, but with heavier tails.

## The F-distribution

If $X \sim \chi^2_n$, and $Y \sim \chi^2_m$ independent of $X$, then

$$U = \frac{X/n}{Y/m} \sim F_{n,m}$$

i.e., the above quantity follows an **F-distribution with** $n$ **numerator and** $m$ **denominator degrees of freedom**.

This distribution plays an important role in hypothesis testing with linear models.