Lecture 1

Distributions and Inference

Aim: to briefly survey the necessary knowledge of random variables, probability distributions and statistical inference.

1. Random variables and probability distributions
2. Statistics of a sample
3. Inferring from a sample
4. Some distribution theory relating to the normal distribution
5. Covariance matrices and some matrix algebra results

Random variables and probability distributions

- Let $Y_1, Y_2, \ldots$ be random variables. Then $y_1, y_2, \ldots$ will denote their observations.
- Y is a **discrete** random variable if it takes a **countable** number of values. For instance, by counting something.
- **Population** P is the set of all possible observations.
- **Probability distribution** is population together with associated probabilities.
- **Probability mass function** describes probability of any given value for a discrete random variable

$$f(y) = \Pr(Y = y)$$

- For a discrete random variable Y we have

$$f(y) \geq 0 \quad \forall y \in P \quad \text{and} \quad \sum_{y \in P} f(y) = 1$$

- Y is a **continuous** random variable it can take on an uncountable set of values. For instance, by measuring something.
- For a continuous random variable Y the $f(y) = \Pr(Y = y)$ is called a **probability density function**, and $\sum_{y \in P} f(y) = 1$ is replaced with $\int_{y \in P} f(y)\, dy = 1$.

- **Cumulative** probability mass function (when P is ordered)

$$F(y) = \Pr(Y \leq y) = \sum_{x \leq y,\, x \in P} f(x)$$

It satisfies

1. $F(-\infty) = 0$
2. $F(\infty) = 1$
3. If $a \leq b$, then $F(a) \leq F(b)$
4. $\Pr(a < Y \leq b) = F(b) - F(a)$

- Example ($P = \{1, \ldots, n\}$, i.e. an n-sided dice)

$$f(y) = \frac{1}{n} \qquad \sum_{y \in P} \frac{1}{n} = 1 \qquad F(y) = \begin{cases} 0 & y \leq 0 \\ y/n & 0 < y < n \\ 1 & y \geq n \end{cases}$$

- The r-th **central moment** of Y:

$$\mathbb{E}(Y^r) = \sum_{y \in P} y^r f(y)$$

$\mu = \mathbb{E}(Y)$ is called the **expected value** or **population mean**.

- The r-th **moment about the mean** of Y

$$\mathbb{E}((Y-\mu)^r) = \sum_{y \in P} (y-\mu)^r f(y)$$

$\sigma^2 = \mathbb{E}((Y-\mu)^2)$ is called **variance** of Y. Its square root σ – **standard deviation**.

Exercise: show that $\sigma^2 = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2$.

- For a continuous random variable Y the sums above should be replaced with the integral $\int_P dy$.

- Let Y_1, Y_2 be discrete random variables taking values in the same population P.
- Their joint probability mass function

$$f(y_1, y_2) = \Pr(Y_1 = y_1, Y_2 = y_2)$$

satisfies

$$\sum_{y_1, y_2 \in P} f(y_1, y_2) = 1$$

- Y_1, Y_2 are independent random variables if

$$f(y_1, y_2) = f_{Y_1}(y_1) f_{Y_2}(y_2)$$

where $f_{Y_1}(y_1)$, $f_{Y_2}(y_2)$ are **marginal** probability mass functions

$$f_{Y_1}(y_1) = \sum_{y_2 \in P} f(y_1, y_2) \qquad f_{Y_2}(y_2) = \sum_{y_1 \in P} f(y_1, y_2)$$

- We define expectation and moments of the joint distributions in a natural way. For a function $g(Y_1, Y_2)$ we define

$$\mathbb{E}(g(Y_1, Y_2)) = \sum_{y_1, y_2 \in P} g(y_1, y_2) f(y_1, y_2)$$

  For example, the expectation value of $g(Y_1, Y_2) = Y_1 Y_2$ is

$$\mathbb{E}(Y_1 Y_2) = \sum_{y_1, y_2 \in P} y_1 y_2 f(y_1, y_2)$$

- The **covariance** between Y_1 and Y_2 is

$$\mathrm{Cov}(Y_1, Y_2) = \mathbb{E}(Y_1 Y_2) - \mathbb{E}(Y_1)\,\mathbb{E}(Y_2)$$

- Random variables Y_1 and Y_2 are **independent** if $\mathrm{Cov}(Y_1, Y_2) = 0$. In this case $\mathbb{E}(Y_1 Y_2) = \mathbb{E}(Y_1)\,\mathbb{E}(Y_2)$.

- For continuous random variables Y_1, Y_2 the sums above should be replaced with the integral $\int\int_P dy_1 dy_2$.

Lecture 1

Distributions and Inference

1. Random variables and probability distributions
2. Statistics of a sample
3. Inferring from a sample
4. Some distribution theory relating to the normal distribution
5. Covariance matrices and some matrix algebra results

Statistics of a sample

- Let $y_1, y_2, \ldots, y_n$ be observations of random variables $Y_1, Y_2, \ldots, Y_n$.

  – Any statistics derived from this sample might differ from those of the underlying population.

  – However, we expect to see systematic patterns induced by the underlying population.

- The sample mean and sample variance are defined by

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

  They are unbiased estimates of population mean μ and variance σ^2.

- Let $Y_1, \ldots, Y_n$ let i.i.d. random variables with mean μ and variance σ^2. Set

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Then

$$\mathbb{E}(\bar{Y}) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(Y_i) = \frac{n}{n}\mu = \mu$$

$$\begin{aligned}\mathrm{Var}(\bar{Y}) &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)\right)^2\right] = \frac{1}{n^2}\,\mathbb{E}\left[\left(\sum_{i=1}^{n}(Y_i - \mu)\right)^2\right] \\ &= \frac{1}{n^2}\,\mathbb{E}\left[\sum_{i=1}^{n}(Y_i - \mu)^2\right] + \frac{1}{n^2}\,\mathbb{E}\left[\sum_{i,j=1,\, i\neq j}^{n}(Y_i - \mu)(Y_j - \mu)\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(Y_i) = \frac{\sigma^2}{n}\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n}(Y_i-\bar{Y})^2 &= \sum_{i=1}^{n}\big(Y_i-\mu-(\bar{Y}-\mu)\big)^2\\ &= \sum_{i=1}^{n}(Y_i-\mu)^2-2\sum_{i=1}^{n}(Y_i-\mu)(\bar{Y}-\mu)+\sum_{i=1}^{n}(\bar{Y}-\mu)^2\\ &= \sum_{i=1}^{n}(Y_i-\mu)^2-2n(\bar{Y}-\mu)^2+n(\bar{Y}-\mu)^2\\ &= \sum_{i=1}^{n}(Y_i-\mu)^2-n(\bar{Y}-\mu)^2\end{aligned}$$

Hence

$$\begin{aligned}\mathbb{E}(S^2) &= \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right)\\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathbb{E}((Y_i-\mu)^2)-n\,\mathbb{E}((\bar{Y}-\mu)^2)\right)\\ &= \frac{1}{n-1}(n\sigma^2-n\sigma^2/n)=\sigma^2\end{aligned}$$

Lecture 1

Distributions and Inference

1. Random variables and probability distributions
2. Statistics of a sample
3. Inferring from a sample
4. Some distribution theory relating to the normal distribution
5. Covariance matrices and some matrix algebra results

Inferring from a sample

- We are usually concerned with making statements about the underlying population by inferring statistical data from a sample.
- We often wish to make an "informed guess" at the value of a population parameter θ using sample data only.

  – We use a certain statistic T to "guess" θ. This statistic is called an **estimator** of θ.

  – The value t of the statistic observed for a particular sample is called an **estimate** of θ.

  – The estimator is called **unbiased** if its estimate coincides with its expectation value, that is if $\mathbb{E}(T) = t$.
- There are three popular methods of estimation: the method of **moments**, **maximal likelihood**, and **least squares**.

The method of moments

- The method of moments chooses the estimate $\hat{\theta}$ of any unknown parameter θ to be that value which makes the population mean μ equal to the sample mean $\bar{y}$.
- In other words, we need to solve the equation

$$\mu(\theta) = \bar{y}$$

  for θ, and name the solution $\hat{\theta}$.
- We illustrate this with an example.

Example. Let $y_1, y_2, \ldots, y_n$ be a sample drawn from a distribution with the probability density function $f(y) = \lambda e^{-\lambda y}$ for $0 \leq y < \infty$.
We want to estimate λ.

– The mean of such a distribution is

$$\mu = \int_0^\infty y\, f(y)\, dy = \int_0^\infty y \lambda e^{-\lambda y} dy = \frac{1}{\lambda}$$

– The sample mean is $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$.

– We identify the mean $\mu = \frac{1}{\lambda}$ with the sample mean $\bar{y}$ giving $\hat{\lambda} = \frac{1}{\bar{y}}$.

– We write $\hat{\lambda}$ to distinguish the estimate from the true parameter λ in question.

- If there are two unknown parameters then we use the first two moments to obtain the estimate. If there are three unknown parameters, we need to use first three moments, and so on.

The method of maximal likelihood

- Let $y_1, y_2, \ldots, y_n$ be a sample drawn from a distribution with a probability density function $f(y;\theta)$ that depends on an unknown parameter θ.

  The **likelihood** of the sample is the joint probability density function

$$L(\theta) = \prod_{i=1}^{n} f(y_i;\theta)$$

- The **maximum likelihood estimator** of θ is a function $\hat{\theta} = g(Y_1, \ldots, Y_n)$ that maximises $L(\theta)$ with respect to θ. Its actual value for a given sample is the **maximum likelihood estimate** (m.l.e.) of θ for that sample.
- The m.l.e. can be interpreted as the value of θ that ascribes the highest possible probability of the sample that was actually obtained.
- The value of θ maximising $L(\theta)$ also maximises $l(\theta) = \log L(\theta)$, and the latter maximisation is often easier to effect in practice.

Example. Let $y_1, y_2, \ldots, y_n$ be a sample drawn from a distribution with probability density function $f(y) = \lambda e^{-\lambda y}$ for $0 \leq y < \infty$.
We want to estimate λ.

– The likelihood and log-likelihood are

$$L(\lambda) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \lambda e^{-\lambda y_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} y_i}$$

$$l(\lambda) = \log(L(\lambda)) = n \log \lambda - \lambda \sum_{i=1}^{n} y_i$$

– Next, we maximize $l(\lambda)$:

$$\frac{\partial l}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} y_i = \frac{n}{\lambda} - n\bar{y} \stackrel{!}{=} 0 \quad \Longrightarrow \quad \hat{\lambda} = 1/\bar{y}$$

– The second derivate test

$$\frac{\partial^2 l}{\partial \lambda^2} = -\frac{n}{\lambda^2} < 0$$

– Hence the m.l.e. of λ is $\hat{\lambda} = 1/\bar{y}$.

The method of least squares

- Let $Y_1, Y_2, \ldots, Y_n$ be such that

$$Y_i = g(\beta_1, \beta_2, \ldots, \beta_k) + \varepsilon_i$$

where $g(\beta_1, \beta_2, \ldots, \beta_k)$ is a function of k scalar parameters $\beta_1, \beta_2, \ldots, \beta_k$ and the ε_i are i.i.d. random variables with zero mean and a common variance σ^2.

- The **least squares estimates** (l.s.e.) of the parameters β_i are the values $\hat{\beta}_i$ which minimize

$$V = \sum_{i=1}^{n} (y_i - g(\beta_1, \beta_2, \ldots, \beta_k))^2$$

for a sample $y_1, y_2, \ldots, y_n$.

- A standard calculus can be employed to find these values. The least squares estimates coincide with with the maximal likelihood ones if the distribution of the ε_i is normal, but not necessarily otherwise.

Example. Let $y_1, y_2, \ldots, y_n$ observations of $Y_1, Y_2, \ldots, Y_n \overset{ind}{\sim} N(\mu, \sigma^2)$ with unknown parameters μ and σ^2. We want to estimate μ.

– Write $Y_i = \mu + \varepsilon_i$ with $\varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$

– We need to minimize $V = \sum_{i=1}^{n} (y_i - \mu)^2$:

$$\frac{\partial V}{\partial \mu} = -2 \sum_{i=1}^{n} (y_i - \mu) = -2(n\bar{y} - n\mu) \overset{!}{=} 0 \quad \Longrightarrow \quad \hat{\mu} = \bar{y}$$

– The second derivative test

$$\frac{\partial^2 V}{\partial \mu^2} = 2n > 0$$

– Thus the l.s.e. of μ is $\hat{\mu} = \bar{y}$.

Lecture 1

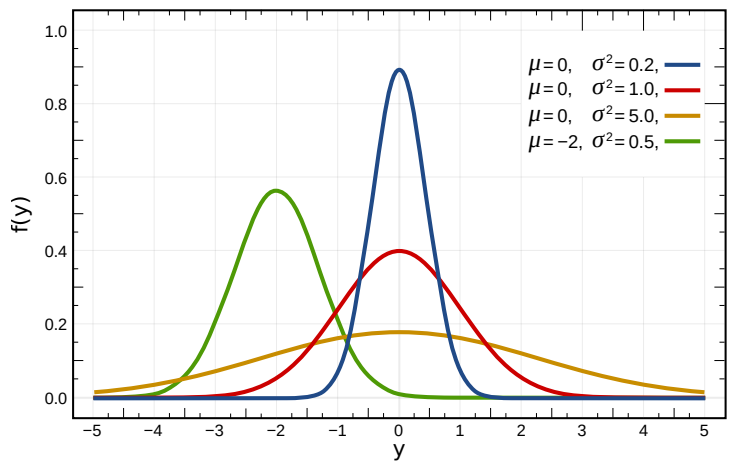
Distributions and Inference

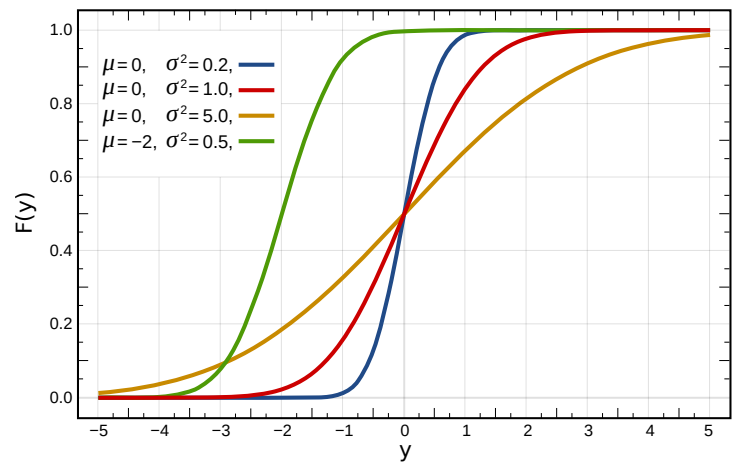
1. Random variables and probability distributions
2. Statistics of a sample
3. Inferring from a sample
4. Some distribution theory relating to the normal distribution
5. Covariance matrices and some matrix algebra results

Some distribution theory relating to the normal distribution

- The normal distribution $N(\mu, \sigma^2)$ is the most used distribution in statisctics. It is parametrized in terms of its mean μ and variance σ^2 only.
- The probability density and cumulative probability distribution functions are

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \qquad F(y) = \int_{-\infty}^{y} f(x)dx$$





- Let $Y \sim N(\mu, \sigma^2)$. Then $Z = (Y - \mu)/\sigma \sim N(0, 1)$ with

$$\varphi(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \qquad \Phi(y) = \int_{-\infty}^{y} \varphi(x)dx$$

- We will need to consider linear combinations of random variables:
  - Let $Y_i \overset{ind}{\sim} N(\mu_i, \sigma_i^2)$ and $a_i \in \mathbb{R}$ for $1 \leq i \leq n$
  - Define a new random variable

$$Z = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n$$

Then

$$\mathbb{E}(Z) = a_1 \mu_1 + a_2 \mu_2 + \cdots + a_n \mu_n$$

$$\text{Var}(Z) = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \cdots + a_n^2 \sigma_n^2$$

Furthermore, Z is normally distributed

$$Z = \sum_{i=1}^{n} a_i Y_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2 \right)$$

Some distribution theory relating to the normal distribution

- We will also need some elements of the multivariate normal distribution:
  - Let $Y_i \sim N(\mu_i, \sigma_i^2)$ for $1 \leq i \leq n$ be normal random variables
  - Set $\sigma_{ij}^2 = \text{Cov}(Y_i, Y_j)$ for $1 \leq i, j \leq n$
- The joint distribution of the Y's is the **multivariate normal distribution**

$$\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, V)$$

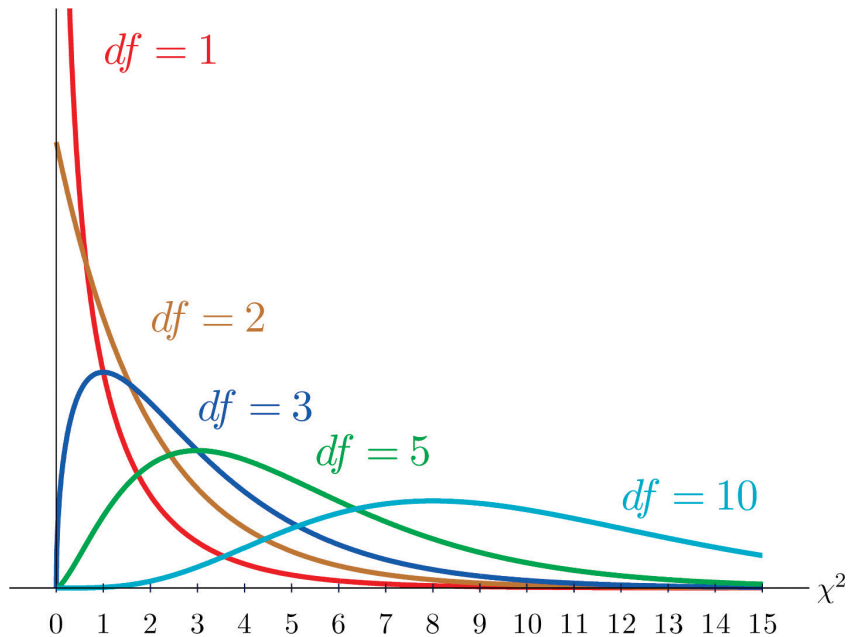
  - $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$
  - mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$
  - variance-covariance matrix V with elements σ_{ij}^2
- You will study multivariate normal distribution in detail in the 3rd year module "Multivariate Statistics".
- There are three distributions which can be derived from the normal distribution and which occur very frequently in all branches of statistics.
  We will introduce them in the next three slides.

Chi-squared distribution

- Let $Y_i \overset{ind}{\sim} N(0,1)$, then

$$Z = Y_1^2 + Y_2^2 + \ldots + Y_n^2 \sim \chi_n^2$$

is distributed according to the **chi-squared** distribution with n d.o.f.

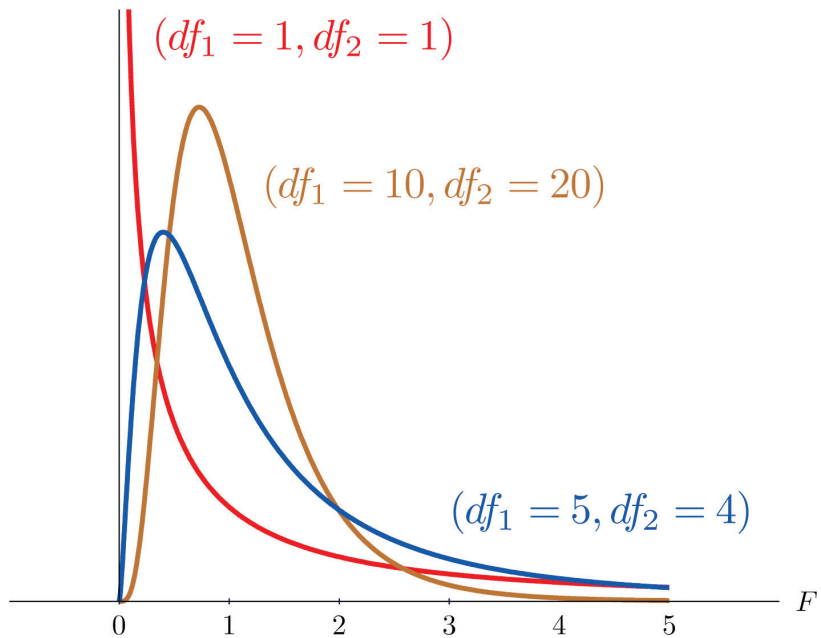


Fisher's F-distribution

- Let $Z \sim \chi^2_p$ and $V \sim \chi^2_q$ be independent random variables, then

$$W = \frac{Z/p}{V/q} \sim F_{p,q}$$

is distributed according to the F-distribution with p and q d.o.f.

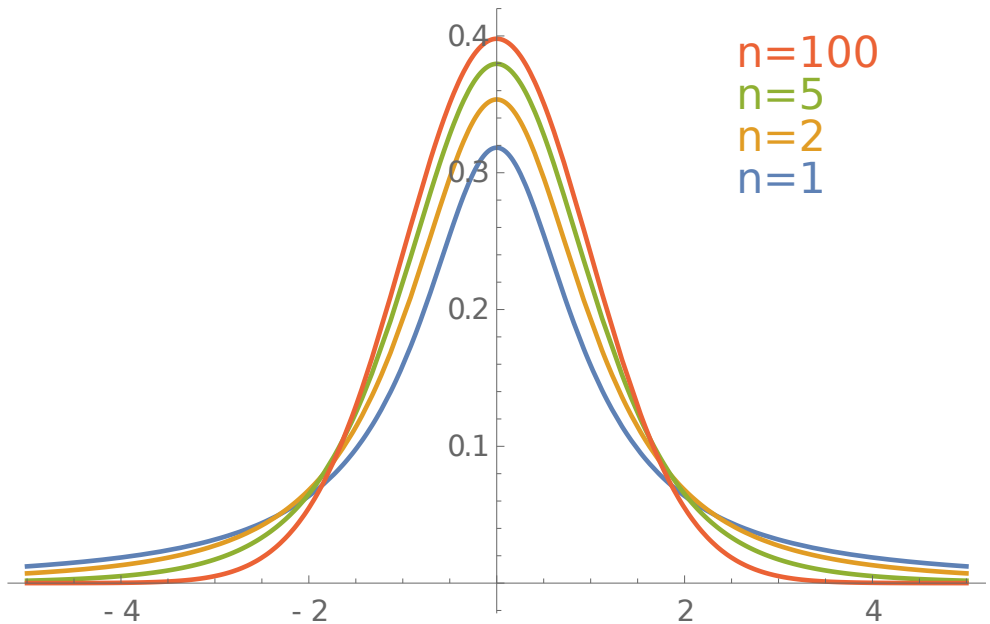


Student's t-distribution

- Let $Y \sim N(0,1)$ and $Z \sim \chi^2_n$ be independent random variables, then

$$W = \frac{Y}{\sqrt{Z/n}} \sim t_n$$

is distributed according to the Student's t-distribution with n d.o.f.



Central Limit Theorem

- Let $\{Y_1, Y_2, \ldots, Y_n\}$ be a sequence of i.i.d. random variables with mean μ and finite variance σ^2 and let $S_n = \sum_{i=1}^{n} Y_i/n$.

- Then as n approaches infinity, the random variable $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $N(0, \sigma^2)$, that is

$$\lim_{n\to\infty} \sqrt{n}(S_n - \mu) \sim N(0, \sigma^2)$$

Lecture 1

Distributions and Inference

1. Random variables and probability distributions
2. Statistics of a sample
3. Inferring from a sample
4. Some distribution theory relating to the normal distribution
5. Covariance matrices and some matrix algebra results

Covariance matrices and some matrix algebra results

- Let $\boldsymbol{z}$ be a random $r \times 1$ vector and let $\boldsymbol{w}$ be a random $p \times 1$ vector.
- The cross covariance of $\boldsymbol{z}$ and $\boldsymbol{w}$ is an $r \times p$ matrix is defined by

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{z}, \boldsymbol{w}) &= \mathbb{E}\big((\boldsymbol{z} - \mathbb{E}(\boldsymbol{z}))(\boldsymbol{w} - \mathbb{E}(\boldsymbol{w}))^T\big) \\
&= \mathbb{E}\big(\boldsymbol{z}\boldsymbol{w}^T - \boldsymbol{z}\,\mathbb{E}(\boldsymbol{w})^T - \mathbb{E}(\boldsymbol{z})\boldsymbol{w}^T + \mathbb{E}(\boldsymbol{z})\,\mathbb{E}(\boldsymbol{w})^T\big) \\
&= \mathbb{E}(\boldsymbol{z}\boldsymbol{w}^T) - \mathbb{E}(\boldsymbol{z})\,\mathbb{E}(\boldsymbol{w})^T - \mathbb{E}(\boldsymbol{z})\,\mathbb{E}(\boldsymbol{w})^T + \mathbb{E}(\boldsymbol{z})\,\mathbb{E}(\boldsymbol{w})^T \\
&= \mathbb{E}(\boldsymbol{z}\boldsymbol{w}^T) - \mathbb{E}(\boldsymbol{z})\,\mathbb{E}(\boldsymbol{w})^T.
\end{aligned}$$

- If $\boldsymbol{z}$ and $\boldsymbol{w}$ are independent variables, $\mathbb{E}(\boldsymbol{z}\boldsymbol{w}^T) = \mathbb{E}(\boldsymbol{z})\,\mathbb{E}(\boldsymbol{w})^T$, their covariance is zero.
- The diagonal entries of the matrix $\mathrm{Cov}(\boldsymbol{z}) = \mathrm{Cov}(\boldsymbol{z}, \boldsymbol{z})$ are the variances of each element of the vector $\boldsymbol{z}$. We will denote the diagonal matrix of variances of $\boldsymbol{z}$ by $\mathrm{Var}(\boldsymbol{z})$.

Covariance matrices and some matrix algebra results

- Assume that $\boldsymbol{z} \sim N_r(\boldsymbol{\mu}, V)$, where $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{z})$ and V is the $r \times r$ variance-covariance matrix. Let $\boldsymbol{c} = \boldsymbol{a} + B\boldsymbol{z}$ for any $p \times 1$ vector $\boldsymbol{a}$ and any $p \times r$ matrix B. Then

$$\boldsymbol{c} \sim N_p\left(\boldsymbol{a} + B\boldsymbol{\mu}, BVB^T\right)$$

- Let M be an $r \times p$ matrix and let N be a $p \times r$ matrix. Then

$$(MN)^T = N^T M^T$$

- The trace of an $r \times r$ square matrix M, written $\mathrm{tr}(M)$ is defined as the sum of its diagonal elements. If S is an $r \times p$ matrix, then

$$\mathrm{tr}(S^T MS) = \mathrm{tr}(MSS^T)$$

- The diagonal part of a matrix M is denoted by $\mathrm{diag}(M)$

Example. Let $Y_1 \sim N(1,2)$ and $Y_2 \sim N(2,3)$ be independent random variables. Find the distribution of $W_1 = Y_1 - Y_2$ and $W_2 = 2Y_1 + 3Y_2$

– Recall

$$Z = \sum_{i=1}^{n} a_i Y_i \sim N\Big(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2 \Big)$$

– Hence

$$\begin{aligned} W_1 &= 1 \cdot Y_1 + (-1) \cdot Y_2 \\ &\sim N(1 \cdot 1 + (-1) \cdot 2,\ 1^2 \cdot 2 + (-1)^2 \cdot 3) = N(-1, 5) \end{aligned}$$

$$\begin{aligned} W_2 &= 2 \cdot Y_1 + 3 \cdot Y_2 \\ &\sim N(2 \cdot 1 + 3 \cdot 2,\ 2^2 \cdot 2 + 3^2 \cdot 3) = N(8, 35) \end{aligned}$$

Example. Let $Y_1 \sim N(0,1)$ and $Y_2 \sim N(2,4)$ be independent random variables. Find the distribution of $W = Y_1^2 + (Y_2 - 2)^2/4$

– Notice that

$$(Y_2 - 2)^2/4 \sim N(0,1)$$

– Hence

$$W = Y_1^2 + (Y_2 - 2)^2/4 \sim \chi_2^2$$

Example. Let $Y_1 \sim N(1,4)$ and $Y_2 \sim N(0,1)$ be independent random variables. Set $\boldsymbol{Y} = \begin{bmatrix} (Y_1 - 1)/2 \\ Y_2 \end{bmatrix}$. Find the distribution of $\boldsymbol{Y}^T \boldsymbol{Y}$

– Notice that

$$(Y_1 - 1)/2 \sim N(0,1)$$

– Hence

$$\boldsymbol{Y}^T \boldsymbol{Y} = ((Y_1 - 1)/2)^2 + Y_2^2 \sim \chi_2^2$$

Next Week

Simple Linear Regression