Lecture 1: Course Introduction & Key Distributions

STA261 – Probability & Statistics II

Ofir Harari

Department of Statistical Sciences

University of Toronto

Outline

Some details for this course

- Text: Rice - *Mathematical Statistics and Data Analysis*
- Coverage: At least Chapters 8, 9 & 13 [skip Bayesian sections; Bootstrap]
- Exposure to some R
- Evaluation: Weekly Quizes [20%], Mid-term [30%], Final [50%]
- Weekly Assignments: Textbook questions on which quiz is based
- Weekly tutorials: Q&A for assigned questions, review of material, quiz
- My office hours: SS 6027, Tuesdays 10:10-11:00, Wednesdays 17:10-18:00 and Fridays 10:10-11:00

Example 1: A Wildlife Study

- Wildlife Population - Caribou
- Population size: N
- Capture D Caribou and tag them
- Release the caribou
- Wait
- Recapture n Caribou
- Count how many are tagged: X
- What's the distribution of X?

- For the population

$$\begin{tikzcd} & N \arrow[dl] \arrow[dr] & \\ N-D & & D \end{tikzcd}$$

not tagged   tagged

- For the recaptured Caribou

$$\begin{tikzcd} & n \arrow[dl] \arrow[dr] & \\ n-X & & X \end{tikzcd}$$

not tagged   tagged

- $X \sim \mathrm{HG}[N, D, n]$

The Hypergeometric distribution

- $X \sim \mathrm{HG}[N, D, n]$
- $P_N(X = x) = \dfrac{\binom{N-D}{n-x}\binom{D}{x}}{\binom{N}{n}}$
- Suppose $N = 100$
- We tag $D = 20$
- Recapture $n = 10$
- Have $X \in \{0, 1, \ldots, 10\}$
- In **R**:
  $P(X = x) = dhyper(x, D, N - D, n)$

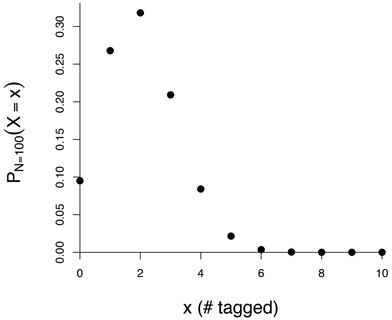
  dhyper(x, D, N-D, n)

```
> dhyper(6,20,80,10)
[1] 0.00354136
> dhyper(c(2,4),20,80,10)
[1] 0.3181706 0.0841073
```

N - population size
D - number of success states
n - number of draws
x - number of observed success

The probability mass function (pmf)

- In STA257 the emphasis was put on the pmf/pdf
- Plot $P_N(X = x)$ –

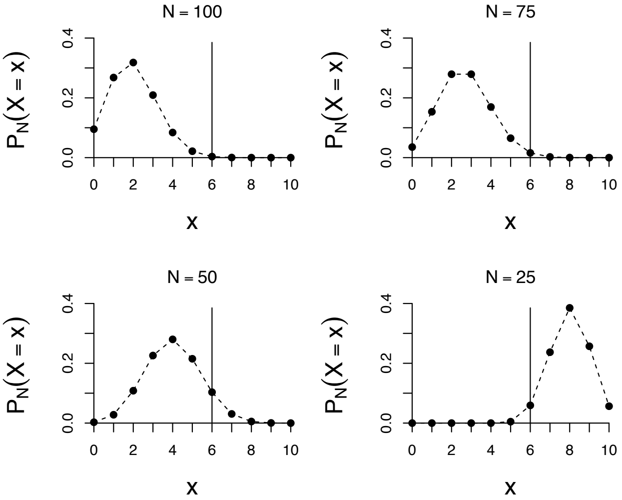



```
> plot(0:10, dhyper(0:10, 20, 80, 10))
```

- In reality the population size N is typically unknown – ecologists use the capture-recapture procedure to *estimate* it.

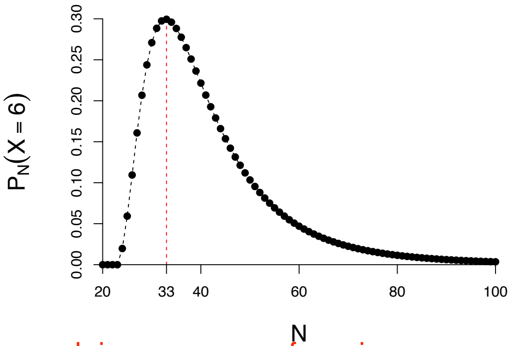
The pmf (continued)

- Suppose $X = 6$ – which N seems the most *likely* to have produced the data?



The Likelihood

- Wait a minute – why are we plotting x vs $P_N(X = x)$ for various N??
- Let's plot N vs $P_N(X = x)$ at $x = x_{\text{obs}} = 6$
- This is the likelihood - $\mathcal{L}(N) = P_N(X = x_{\text{obs}})$



this graph is a summary of previous ones

```
> plot(20:100, dhyper(6, 20, 0:80, 10))
```

Maximum Likelihood Estimation

- The likelihood function obtained its maximum at $N = 33$
  - but what does this mean?
- Intuitively, a good choice for an estimate of the population size
- Is it a good estimate though?
  - Maybe provide a range of plausible values for N?
- Other methods of estimation?
- The topic of Parameter Estimation will be our focus for the first half of this course.

parameter estimation — mean, variance…

Example 2: Evidence of discrimination

- 48 Files: 24 women, 24 men
- Randomly assigned to 48 male supervisors
- Assessed as *promote* or *hold*
- However all files are identical – just labeled male or female

	Male	Female	
Promote	21	14	35
Hold	3	10	13
	24	24	48

- 21 Men promoted versus 14 women
- Is there a bias against women?
- Is there any evidence of bias?
- Could 21 & 14 happen by chance?

Evidence of discrimination (continued)

- Could 21 & 14 happen by chance?
- Sure! In fact, so could 24 & 0
- But *unlikely* in the absence of bias
- But is 21 & 14 unlikely?
- We need to quantify this
- Assume there's no bias, and calculate the probability of observing 21 & 14

- We have a pool of 48 files (balls)
- 24 of which are males (white balls) and the rest are females (black)
- No bias means that 35 promoters (balls) were sampled **at random** with no replacement – and 21 of them turned out to be males (white).
- What's the *probability* of this?
- This is *Statistical Inference*

Statistical Inference

- Let X be the number of *promote* files of type *male*
- Based on our above argument, if there's no bias $X \sim \mathrm{HG}(48, 24, 35)$ (48 balls, 24 white, 35 sampled)
- What is the probability of promoting at least as many males then (under the "no bias" assumption)? $P(X \geq 21)$
-
```
> phyper(20, 24, 24, 35, lower.tail = FALSE)
[1] 0.02449571
```
- If we carried out this experiment repeatedly, about 1 out of every 41 times would yield results as extreme as these (21 & 14). This is considered rather rare – strong evidence for bias!
- This is how science works: when proposing a new theory, one must show that the existing theory is unlikely to reproduce the observed data
- This is Hypothesis Testing – the focus of the 2nd half of this course.

Probability vs. Statistics

- Parameter θ of scientific interest
- Data X from a process modelled using θ
- Probability:
  - ⋆ $\theta \longrightarrow$ Fixed and known
  - ⋆ $X \longrightarrow$ Random and unknown
- Straightforward [no inference]
- Statistical Inference:
  - ⋆ $\theta \longrightarrow$ Unknown
  - ⋆ $X \longrightarrow$ Observed: Fixed and known
- Inference required through [statistical/quantitative] reasoning/thinking

UNIVERSITY OF TORONTO

Basic concepts

Definition

The random variables $X_1, \ldots, X_n$ are called a *random sample of size n from the population* $f(x)$ if $X_1, \ldots, X_n$ are all independent r.v.'s and the marginal p.d.f (in the continuous case) or p.m.f (in the discrete case) of each X_i is $f(x)$.

- Alternatively, we may just write $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(x)$ to denote that $X_1, \ldots, X_n$ are independent and identically distributed.
- I will use these two forms interchangeably throughout the course

Definition

Any function $g(X_1, \ldots, X_n)$ of the sample is called a (sample) *statistic*.

UNIVERSITY OF TORONTO

Basic concepts (continued)

- Any summary of the data is a statistic –
  - the first observation
  - the sample maximum
  - the first quartile
  - etc.

  (but some statistics are obviously of greater interest than others)

- Statistics, as functions of random variables, are themselves random variables.

Definition

The distribution of a statistic $T = g(X_1, \ldots, X_n)$ is called the *sampling distribution* of T.

Example: the sample mean

Perhaps the most famous sample statistic is the sample mean.

Definition

Let $X_1, \ldots, X_n$ be a random sample from some distribution/population. The sample mean is

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Suppose that the distribution of the X_i's has mean μ and variance σ^2. Then, from Probability theory –

- $\mathbb{E}\left[\overline{X}\right] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[X_i\right] = \frac{1}{n} \cdot n\mu = \mu.$
- $\mathrm{Var}\left[\overline{X}\right] = \mathrm{Var}\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}\left[X_i\right] = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}.$
- Fine, we know the mean and the variance of $\overline{X}$, but is that a complete characterization of its sampling distribution? Sort of...

The sample mean (continued)

- In general, the distribution of $\overline{X}$ depends on the distribution of the X_i's.

proved mgf for sample mean is normal

- If the X_i's follow a $\mathcal{N}(\mu, \sigma^2)$ then $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.
- If n is "large", then even when the X_i's themselves are not normal, the Central Limit Theorem applies and $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ (at least approximately), in the sense that

$$P(\overline{X} \leq t) \approx \Phi\left(\frac{t-\mu}{\sigma/\sqrt{n}}\right).$$

used in computation

The χ^2 distribution

Definition

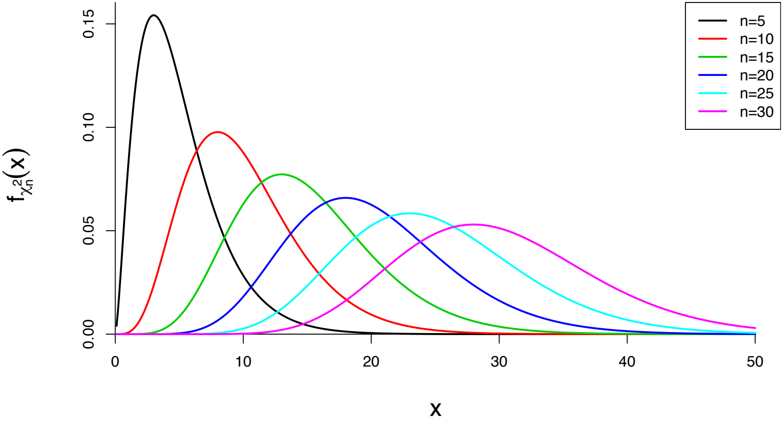
Let $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. The distribution of the statistic $X^2 = \sum_{i=1}^{n} Z_i^2$ is called the chi-square distribution with n degrees of freedom and is denoted by χ_n^2.

Some facts about the χ_n^2 distribution:

- It is a special case of the Gamma distribution: $\chi_n^2 = \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$. Consequently, $Y \sim \chi_n^2$ has $\mathbb{E}[Y] = n$ and $\mathrm{Var}[Y] = 2n$.
- The *Moment Generating Function* (mgf) of $Y \sim \chi_n^2$ is $M_Y(t) = (1-2t)^{-n/2}$.
- If $X \sim \chi_m^2$ and $Y \sim \chi_n^2$ are independent r.v.'s then $X + Y \sim \chi_{m+n}^2$ (easily verifiable through application of the mgf).

What does it look like?

The χ^2 pdf



- **All well and good, but when do we ever encounter the χ^2 distribution?**

Example: the sample variance

Definition

The sample variance of a random sample $X_1, \dots, X_n$ is defined to be

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

- Seems intuitive, recalling that the (infinite) population variance is given by $\mathrm{Var}[X] = \mathbb{E}\left\{(X - \mathbb{E}[X])^2\right\}$.
- The $n-1$ in the denominator seems odd though – why not n?
  - To be clarified shortly.
- What can we say about the sampling distribution of S^2?
  - The answer to this requires additional assumptions and some more effort on our part.

The connection between S^2 and the χ^2 distribution

Theorem

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, and let S^2 be the sample variance. Then

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

- Note that $\dfrac{(n-1)S^2}{\sigma^2} = \displaystyle\sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma}\right)^2$

replacement informal: notice mu here!

- Also note that $\displaystyle\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n} Z_i^2 \sim \chi^2_n$
- It appears that replacing μ by the "next best thing" $\overline{X}$ cost us one degree of freedom.
- To prove the above Theorem, we will need the following powerful result.

χ^2 and the sample variance (continued)

Proposition

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, and let $\overline{X}$ and S^2 be the sample mean and variance, respectively. Then $\overline{X}$ and S^2 are independent.

- I find the proof appearing in the book lacking. A more complete (and quite involved) proof has been uploaded to Blackboard.
- All is now set to prove the Theorem.

Proof:

First, it should come as no surprise that

$$\sum_{i=1}^{n}(X_i - \bar{X}) = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \bar{X} = n\overline{X} - n\overline{X} = 0.$$

χ^2 and the sample variance (continued)

Proof (continued):

Now,

$$\sum_{i=1}^{n}(X_i-\mu)^2 = \sum_{i=1}^{n}(X_i-\overline{X}+\overline{X}-\mu)^2$$

$$= \sum_{i=1}^{n}(X_i-\bar{X})^2 + \sum_{i=1}^{n}(\bar{X}-\mu)^2 + 2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\overline{X})$$

$$= \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu)^2,$$

or

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 + \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2.$$

χ^2 and the sample variance (continued)

Proof (continued):

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2=\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2+\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$$

- The LHS is the sum of n independent squared standard normals
  - Follows a χ^2_n distribution, with mgf $M_U(t)=(1-2t)^{-n/2}$
- The first term on the RHS is $V=\dfrac{(n-1)S^2}{\sigma^2}$ with mgf $M_V(t)$
- The second term on the RHS is a squared standard normal
  - Follows a χ^2_1 distribution, with mgf $M_W(t)=(1-2t)^{-1/2}$
- The two terms on the RHS are independent, hence

specifically independence of \bar{X} and S^2

$$M_U(t)=M_V(t)\cdot M_W(t)\Longrightarrow M_V(t)=\frac{M_U(t)}{M_W(t)}=(1-2t)^{-\frac{n-1}{2}}$$

and use M_{V+W} = M_V X M_W

χ^2 and the sample variance (continued)

Proof (continued):

- We have shown that $V = \dfrac{(n-1)S^2}{\sigma^2}$ has the mgf $M_V(t) = (1-2t)^{-\frac{n-1}{2}}$
- But this is the mgf of a χ^n_{n-1} random variable
- From the uniqueness of the mgf, $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^n_{n-1}$.

Student's t distribution

Definition

The distribution of the quotient $\dfrac{Z}{\sqrt{U/\nu}}$, where $Z \sim \mathcal{N}(0,1)$, $U \sim \chi^2_\nu$ and Z and U are independent is called *Student's t distribution with ν degrees of freedom* and is denoted by t_ν.

Tidbit: British statistician William Gosset worked at the Guinness Brewery in Dublin, Ireland. He ran into the t distribution while working on problems involving small samples, and published a paper using the pseudonym "Student" to hide his (and his employer's) identity.

What gives rise to Student's t distribution?

Example

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, and let $\overline{X}$ and S^2 be the sample mean and variance, respectively. Then $T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$.

this is a standardization process for sample mean\

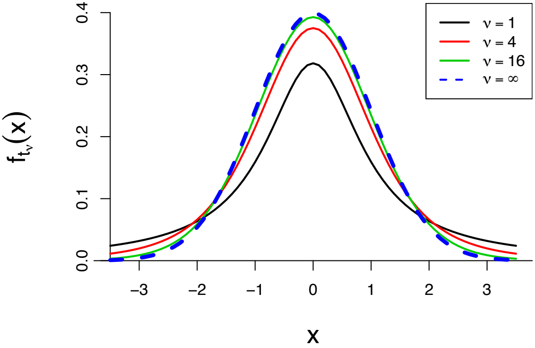
- Curiously, on replacing σ with S in the "standardization" of $\overline{X}$, we lost a degree of freedom again.

Proof:

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \bigg/ \frac{S}{\sigma} = \frac{\mathcal{N}(0,1)}{\dfrac{1}{\sqrt{n-1}} \cdot \sqrt{\dfrac{(n-1)S^2}{\sigma^2}}} = \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}}.$$

Because $\overline{X}$ and S^2 are independent, so are the numerator and the denominator, hence $T \sim t_{n-1}$.

What does Student's t distribution look like?



- Symmetric about 0
- "Heavier" tails than the standard normal distribution
- $t_\infty = \mathcal{N}(0, 1)$

The F distribution

Definition

The distribution of the quotient $\dfrac{U/m}{V/n}$, where $U \sim \chi^2_m$, $V \sim \chi^2_n$ and U and V are independent is called the F *distribution with* m *and* n *degrees of freedom* and is denoted by $F_{m,n}$.

The F distribution is named in honor of Sir Ronald Fisher (1890-1962), the father of modern-time statistics.

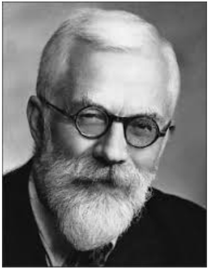


Figure: Source: Wiley Online Library

When do we encounter the F distribution?

Example

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be two (independent) random samples from a $\mathcal{N}(\mu, \sigma^2)$ distribution, and let S_X^2 and S_Y^2 be their sample variances, respectively. Then $F = S_X^2 / S_Y^2 \sim F_{m-1,n-1}$.

Proof: try to make chi-squared from s^2, use the formula

$$F = \frac{S_X^2}{S_Y^2} = \frac{(m-1)S_X^2}{(m-1)\sigma^2} \Bigg/ \frac{(n-1)S_Y^2}{(n-1)\sigma^2} = \frac{\chi^2_{m-1}/(m-1)}{\chi^2_{n-1}/(n-1)}.$$

- Because the two samples are independent, so are S_X^2 and S_Y^2, and $F \sim F_{m-1,n-1}$.

How to do calculations

- Toward the end of the book there are χ^2, t and F tables
  - ⋆ Pretty old fashioned, but will be used during the exam – better get used to them
- In **R**:

```
> dchisq(3.2, 12) #chi-square pdf with 12 d.f. evaluated at x=3.2
[1] 0.008820993
> pchisq(1.15, 9) #chi-square cdf with 9 d.f. evaluated at x=1.15
[1] 0.0009931852
> qchisq(0.95, 6) #The 95% quantile of a chi-square dist. with 6 d.f.
[1] 12.59159
> rchisq(3, 23) #drawing 3 random numbers from a chi-square dist. with 23 d.f.
[1] 23.89533 28.01067 18.87753
```