Session 1

Probability & Statistics Review

Anthony Tay

This Version: 20 Aug 2025

Course Objectives

- Theoretical foundations of OLS estimation of linear regression model
  - when it works well and when it doesn't
  - application to causal analysis / hypothesis testing / prediction
  - using matrix algebra
- Extend OLS to IV and GLS Estimation
- Introduction to more advanced topics
  - Time series
  - IV and GMM estimation
  - Panel data models
  - MLE estimation and limited dependent variable models

Session 1

- A Bit of Math
- Probability Review: Random variable, pdfs, mean, variance
- Statistics Review: Estimation, Hypothesis Testing
- Course Administrative Arrangements
  - Course webpage vs Course eLearn page, Grading, Assignments

Session 1 Introduction

What is **statistics**?

How to learn about a **population** given a **sample** of observations from that population.

- US non-institutional working civilians aged 16 or above in 2018. Interested in average hourly earnings in this population?
- All Singapore households in 2020. Interested in no. of dogs per SG household on average?

Session 1 Introduction

- All goods that a machine can potentially produce in its lifetime. Interested in the defect rate?
- Outcomes of an infinite number of potential tosses of a coin. Interested in whether coin is fair
- Price and output of an economy over time. Interested in relationship between output growth and inflation?

Last three better described as **processes**, the population of interest being the set of all potential outcomes of the process

Probability theory is the mathematics of chance. Statistics uses probability theory to develop and evaluate estimators for prediction, hypothesis testing, causal analysis, etc.)

Math Review: Summation Notation

Given a set of numbers $\{x_i\}_{i=1}^n = \{x_1, x_2, ..., x_n\}$, define

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + ... + x_n$$

Two Rules:

- $\sum_{i=1}^n (a_i + b_i) = \sum_{i=1}^n a_i + \sum_{i=1}^n b_i$
- $\sum_{i=1}^n ca_i = c\sum_{i=1}^n a_i$ where c is some constant value

Summation Notation

Two Results: For any set of numbers $\{x_i, y_i\}_{i=1}^n$ we have

- Sum of deviations from sample mean is zero

$$\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \overline{x} = n\overline{x} - n\overline{x} = 0\,, \quad \text{where } \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

```
x <- c(1, 4, 2, pi, exp(1), 100000)  # insert whatever numbers you want
sum(x - mean(x))
```

```
[1] -3.637979e-12
```

Summation Notation

- Sum of product of deviation from sample means (alternative expressions)

$$\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})=\sum_{i=1}^{n}(x_i-\overline{x})y_i=\sum_{i=1}^{n}x_i(y_i-\overline{y})=\sum_{i=1}^{n}x_iy_i-n\overline{x}\,\overline{y}$$

```
x <- c(1, 4, 2, pi, exp(1), 1000)             # insert whatever numbers you want
y <- c(5, 3029, 2911, sin(4.32), 1.43, 403)   # insert whatever numbers you want
c(sum((x - mean(x))*(y-mean(y))), sum((x-mean(x))*y), sum(x*(y-mean(y))),
  sum(x*y) - length(x)*mean(x)*mean(y))
```

```
[1] -650747.2 -650747.2 -650747.2 -650747.2
```

Summation Notation

Proof of first equality

$$\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) = \sum_{i=1}^{n}(x_i-\overline{x})y_i - \sum_{i=1}^{n}(x_i-\overline{x})\overline{y}$$

$$= \sum_{i=1}^{n}(x_i-\overline{x})y_i - \overline{y}\underbrace{\sum_{i=1}^{n}(x_i-\overline{x})}_{=0}$$

$$= \sum_{i=1}^{n}(x_i-\overline{x})y_i$$

Probability Review

Random variable: numerical outcome X of an action where

- there is a range of possible outcomes
- there is randomness in terms of which outcome is obtained each time the action is taken

Examples

- draw from population of working citizens, observe average working hours in previous year, or draw from population of households, observe number of dogs
- toss a coin, observes heads or tails (recode to 1 and 0)
- everyday actions and decisions of individuals and firms in an economy, observe output and prices

Probability Review

Random does not mean arbitrary

- Outcomes that occur follow some "rules" regarding relative chance of occurrence
- Summarized in **probability distribution function (pdf)**
- We often use pdf as a "model" of a population

E.g. $X \sim$ Bernoulli(p)

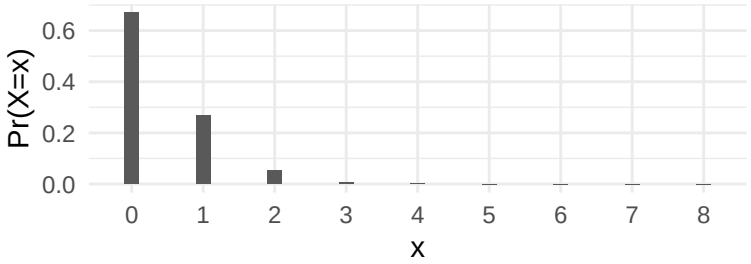
- Possible outcomes $x = 0, 1$ with $\Pr(X = 0) = 1 - p$ and $\Pr(X = 1) = p$
- $\Pr(X = x)$ where $x = 0, 1$ and

$$f_X(x) = \Pr(X = x) = p^x (1 - p)^{1-x}\,,\ x = 0, 1\,.$$

Probability Review

E.g., $X \sim \text{Poisson}(\lambda)$

$$f_X(x) = \Pr(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \; x = 0, 1, 2, \ldots .$$



Probability Review

Bernoulli: discrete random variable with finite range

Poisson: discrete random variable with countably infinite range

Next example: continuous random variable

- Discrete random variable pdf $f_X(x) = \Pr(X = x)$
- Continuous random variable pdf $f_X(x)$ st $\Pr(a \leq X \leq b) = \int_a^b f_X(x)\,dx$.

For discrete rv: **probability mass function**

For continuous rv: **probability density function**

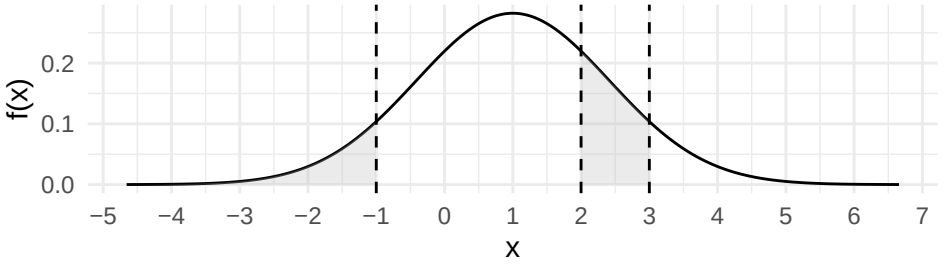
We will call both **probability distribution functions**

Probability Review

E.g. $X \sim \text{Normal}(\mu, \sigma^2)$, Standard Normal if $\mu = 0$ and $\sigma^2 = 1$:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \; x \in \mathbb{R}.$$

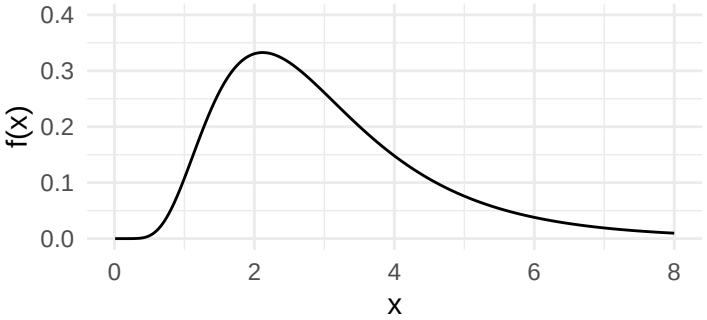
Normal$(1, 2)$ pdf with $\Pr(X \leq -1)$ and $\Pr(2 \leq X \leq 3)$ shaded:



Probability Review

$X \sim \text{Log-normal}(\mu, \sigma^2)$

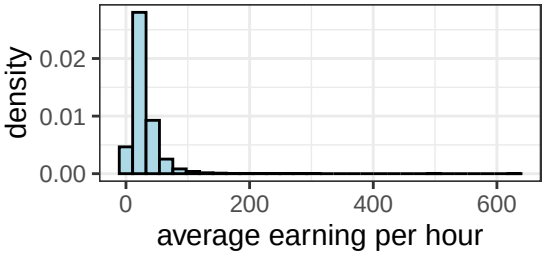
$$X \sim \text{Log-normal}(\mu, \sigma^2) \iff \ln X \sim \text{Normal}(\mu, \sigma^2)$$

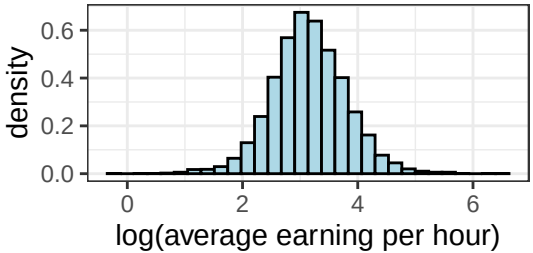


Probability Review

Data set `earnings2019.csv` contains a random sample of almost 5000 U.S. non-institutional civilians 16 yrs and above who worked in 2018 (these individuals are part of the 2019 wave of the University of Michigan Panel Survey of Income Dynamics).

Histogram estimates of $earn$ and $\ln earn$





Probability Review: Expected Values

Mean or **expected value** of a random variable X is defined as

$$E(X) = \begin{cases} \sum_x x f_X(x) = \sum_x x \Pr(X = x) & \text{if } X \text{ is discrete, and} \\ \int_{-\infty}^{+\infty} x f_X(x)\, dx & \text{if } X \text{ is continuous.} \end{cases}$$

Measures "center" or "location" of distribution

- If $X \sim \text{Bernoulli}(p)$, then $E(X) = 1 \cdot p + 0 \cdot (1 - p) = p$.
- If $X \sim \text{Poisson}(\lambda)$, then

$$E(X) = \sum_{x=0}^{\infty} x \Pr(X = x) = \sum_{x=0}^{\infty} \frac{x e^{-\lambda} \lambda^x}{x!} = \lambda\,.$$

Probability Review: Expected Values

- If $X \sim \text{Normal}(\mu, \sigma^2)$, then

$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx = \mu\,.$$

- If $X \sim \text{Log-normal}(\mu, \sigma^2)$, then

$$E(X) = \exp(\mu) \exp\left(\frac{\sigma^2}{2}\right)$$

Alternative measure of location is median: $\Pr(X \leq \text{Median}) = 0.5$

- If $X \sim \text{Log-normal}(\mu, \sigma^2)$, then $Median(X) = \exp(\mu)$

Probability Review: Expected Values

If X is a random variable, then $g(X)$ is also a random variable, with expectation:

$$E(g(X)) = \begin{cases} \sum_X g(x) f_X(x) = \sum_X g(x) \Pr(X = x) & \text{if } X \text{ is discrete, and} \\ \int_{-\infty}^{+\infty} g(x) f_X(x)\, dx & \text{if } X \text{ is continuous.} \end{cases}$$

For example,

- if $X \sim$ Bernoulli(p), we have $E(X^2) = 1^2 p + 0^2(1-p) = p$
- if $X \sim$ Poisson(λ), then $E(X^2) = \lambda + \lambda^2$
- if $X \sim$ Normal(μ, σ^2), then $E(X^2) = \sigma^2 + \mu^2$.

Probability Review: Expected Values

More properties:

- $E(ag(X) + bh(X)) = aE(g(X)) + bE(h(X))$ where a and b are constants.
- $E(a) = a$ for constants a
- $E(a + bX) = a + bE(X)$
- if X and Y are any two random variables, then

$$E(aX + bY) = aE(X) + bE(Y)\,.$$

Probability Review: Variance

The **variance** of a random variable X is its expected squared deviation from mean, i.e.,

$$Var(X) = E((X - E(X))^2) = \begin{cases} \sum_X (x - E(X)^2) f_X(x) & \text{if } X \text{ is discrete, and} \\ \int_{-\infty}^{+\infty} (x - E(X))^2 f_X(x)\, dx & \text{if } X \text{ is continuous.} \end{cases}$$

- Measures "spread" of a probability distribution
- Square root of $Var(X)$ is the **standard deviation** of X
- Unit of measurement of the standard deviation follows that of the variable itself

Another expression for the variance:

$$Var(X) = E((X - E(X))^2) = E(X^2 - 2XE(X) + E(X)^2) = E(X^2) - E(X)^2$$

Probability Review: Variance

- If $X \sim$ Bernoulli(p), then $Var(X) = p(1-p)$
- If $X \sim$ Poisson(λ), then $Var(X) = \lambda$
- If $X \sim$ Normal(μ, σ^2), then $Var(X) = \sigma^2$

Other properties:

- $Var(aX + b) = a^2\, Var(X)$
- $Var(aX + bY) = a^2\, Var(X) + b^2\, Var(Y) + 2ab\, Cov(X, Y)$

Covariance covered in detail later. We note, for the moment, that if two random variables X and Y are **independent** (meaning that knowing the outcome of one tells you nothing about the other), then $Cov(X, Y) = 0$, and

$$Var(aX + bY) = a^2\, Var(X) + b^2\, Var(Y) \qquad \textit{NB: Only if } Cov(X, Y) = 0.$$

Probability Review: Jensen's Inequality

Since

$$Var(X) = E(X^2) - E(X)^2$$

and $Var(X) \geq 0$, we have

$$E(X^2) \geq E(X)^2$$

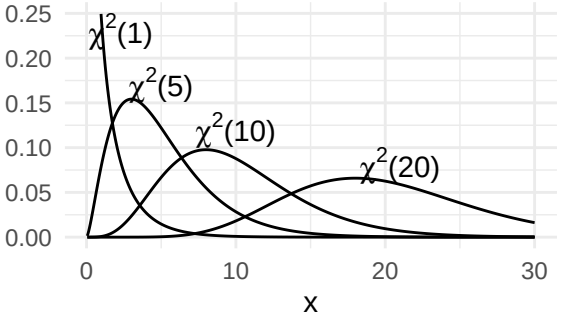
Special case of **Jensen's inequality**: for any random variable X,

$$E(g(X)) \geq g(E(X)) \quad \text{for any convex function } \ g\,.$$

The inequality is reversed if g is concave. Equality holds if $g(X) = a + bX$.

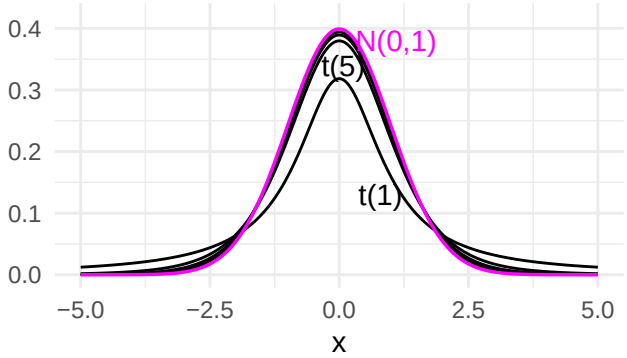
Probability Review: Chi-Square Distribution

- If $X \sim \text{Normal}(0,1)$, then $X^2 \sim \chi^2(1)$.
- If $X_1, X_2, \ldots, X_k$ are independent Normal$(0,1)$, then $\sum_{i=1}^{k} X_i^2 \sim \chi^2(k)$.
- If $X \sim \chi^2(k)$, then $E(X) = k$ and $Var(X) = 2k$



Probability Review: Student t-Distribution

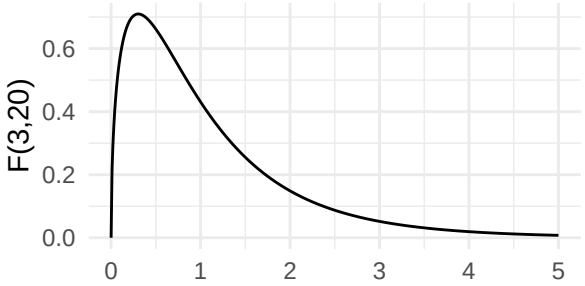
If $X \sim \text{Normal}(0,1)$ and $W \sim \chi^2(v)$ and independent, then $\dfrac{X}{\sqrt{W/v}} \sim \text{t}(v)$



Probability Review: F Distribution

If $W_1 \sim \chi^2(v_1)$ and $W_2 \sim \chi^2(v_2)$ independent, then $\dfrac{W_1/v_1}{W_2/v_2} \sim F(v_1, v_2)$

- If $X \sim t(v)$, then $X^2 \sim F(1, v)$,
- If $X \sim F(v_1, v_2)$, then the pdf of $v_1 X$ converges to the $\chi^2(v_1)$ pdf as $v_2 \to \infty$.



Statistics: Estimation

Statistics: Learning about a certain population using information from a (possibly small) sample from that population

Example of a Population: Non-institutional employed civilians aged 16 and above in US in 2018

Population Characteristics of Interest:

- Average Hourly Earnings
- Relationship between Hourly Earnings and Other Variables (Next week)

Random Sample

Random Sample

- Every individual in population has equal chance of getting selected (so sample "looks like" the population)
- One individual sampled does not result in another being more or less likely to be sampled

Data in `earnings2019.csv`

- Collected by U. Michigan's Institute for Social Research as part of their 2019 wave of their Panel Study of Income Dynamics
- N = 4946 individuals after filtering for employment (defined as ≥ 1000 hrs worked in 2018)

Data Example

```
dat <- read_csv("data\\earnings2019.csv", show_col_types=FALSE)
head(dat,3)
```

```
# A tibble: 3 x 11
    age height  educ feduc meduc tenure  wexp race   male  earn totalwork
  <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <chr> <dbl> <dbl>     <dbl>
1    59     67    12     3     3      5    30 White     0 36.3       1652
2    43     63    10     4     3      7    13 White     1  6.46      1548
3    28     74    12     2     3      6     9 White     1 13.1       2460
```

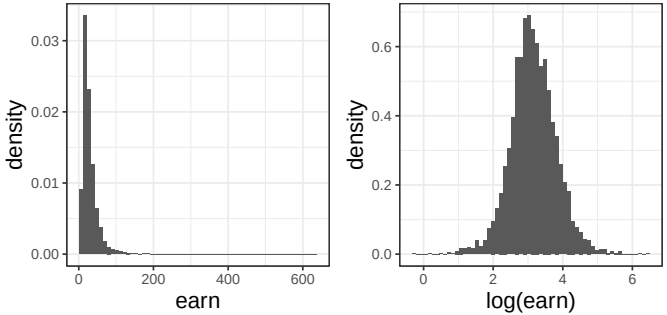
Data Example (Summary of Selected Variables)

```
dat %>% select(-c(race, feduc, meduc)) %>% summary(dat)
```

```
      age            height           educ           tenure      
 Min.   :19.00   Min.   :40.00   Min.   : 7.00   Min.   : 1.000  
 1st Qu.:33.00   1st Qu.:64.00   1st Qu.:12.00   1st Qu.: 3.000  
 Median :40.00   Median :67.00   Median :14.00   Median : 6.000  
 Mean   :41.99   Mean   :67.45   Mean   :14.31   Mean   : 9.177  
 3rd Qu.:51.00   3rd Qu.:70.00   3rd Qu.:16.00   3rd Qu.:13.000  
 Max.   :82.00   Max.   :83.00   Max.   :17.00   Max.   :54.000  
      wexp             male             earn            totalwork    
 Min.   : 1.000   Min.   :0.0000   Min.   :  0.7428   Min.   :1000  
 1st Qu.: 3.000   1st Qu.:0.0000   1st Qu.: 15.5048   1st Qu.:1936  
 Median : 7.000   Median :0.0000   Median : 22.9995   Median :2080  
 Mean   : 9.251   Mean   :0.4646   Mean   : 29.2315   Mean   :2182  
 3rd Qu.:13.000   3rd Qu.:1.0000   3rd Qu.: 35.0235   3rd Qu.:2428  
 Max.   :51.000   Max.   :1.0000   Max.   :628.9308   Max.   :5824  
```

Data Example (Distribution of earn and ln earn)

```
plot_theme <- theme_bw() + theme(aspect.ratio = 1, axis.title = element_text(size=14))
p1 <- ggplot(dat, aes(x=earn)) + geom_histogram(aes(y=after_stat(density)), bins=60, boundary=0) +
      scale_x_continuous("earn") + plot_theme
p2 <- ggplot(dat, aes(x=log(earn))) + geom_histogram(aes(y=after_stat(density)), bins=60, boundary=0) +
      scale_x_continuous("log(earn)") + plot_theme
p1 | p2
```



Statistical Model

Statistical Model: A stylized description of the population and how you obtained sample $\{X_i\}_{i=1}^n$

E.g. A statistical model

$$X_i \overset{iid}{\sim} \text{Log-normal}(\mu, \sigma^2)$$

where

- X_i is $earn$ for individual i
- "iid" stands for independently and identically distributed (another interpretation of "random sample")

Statistical Model

Often we **do not** need to specify distribution fully

We can assume

$$X_i \text{ iid such that } E(X_i) = \mu \text{ and } Var(X_i) = \sigma^2 < \infty \text{ for all } i = 1, \ldots, n\,.$$

NB: μ and σ^2 here is used as "generic notation" for mean and variance. E.g., if X_i is bernoulli, then $\mu = p$, $\sigma^2 = p(1-p)$. If X_i is Poisson, then $\mu = \sigma^2 = \lambda$, etc.

Very general model! Assumes only that:

- sample is a random sample
- population is well-represented by *some* distribution with a finite mean and a finite variance

Statistical Estimators

We want to estimate μ (population mean)

Since $\mu = E(X)$, it seems reasonable to estimate μ using:

- $\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ "Sample Mean"

Is this a good idea?

We need to define what makes a "good estimator"

Bias

One commonly used criterion is **unbiasedness**: $E(\hat{\theta}) = \theta$

Sample mean is unbiased for true mean (under our stated conditions):

Proof: $E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) = \frac{1}{n} n\mu = \mu$

- You will not *systematically* over- or under-estimate the population mean.
- (Thought experiment) If, say, 200 people went to the population and each obtained a random sample of n individuals and calculated the sample mean. Each would obtain a different sample mean, but their sample means will be nicely centered around the true (unknown) population mean.

Standard Error of Sample Mean

We should always try to get some idea of the potential size of estimation error

Consider variance of sample mean

$$Var(\overline{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

- Not operational since σ^2 is unknown
- But already informative – variance goes down with increasing sample size

Standard Error of Sample Mean

To get numerical estimate of potential estimation error size:

- estimate σ^2 (let's call it $\widehat{\sigma^2}$ for now)
- Use square root of $\widehat{Var}(\overline{X}) = \dfrac{\widehat{\sigma^2}}{n}$, i.e., s.e.$(\overline{X}) = \sqrt{\dfrac{\widehat{\sigma^2}}{n}}$

Since $Var(X) = \sigma^2 = E((X - E(X))^2) = E(X^2) - E(X)^2$, it seems reasonable to use

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

Unfortunately, this is a (downward) biased estimator for σ^2

Standard Error of Sample Mean

Proof:

- Since $Var(X_i) = E(X_i^2) - E(X_i)^2$, we have $E(X_i^2) = \sigma^2 + \mu^2$
- Since $Var(\overline{X}) = E(\overline{X}^2) - E(\overline{X})^2$, and $E(\overline{X}) = \mu$, we have $E(\overline{X}^2) = Var(\overline{X}) + \mu^2$
- Furthermore, we have $Var(\overline{X}) = \dfrac{\sigma^2}{n}$

Therefore:

$$E\left(\widetilde{\sigma^2}\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E(\overline{X}^2) = \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \frac{n-1}{n}\sigma^2$$

Standard Error of Sample Mean

Fortunately, in this case, there is an obvious unbiased estimator. Let

$$\widehat{\sigma^2} = \frac{n}{n-1}\widetilde{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \qquad \text{(sample variance)}$$

Then

$$E(\widehat{\sigma^2}) = \left(\frac{n}{n-1}\widetilde{\sigma^2}\right) = \frac{n}{n-1}\left(\widetilde{\sigma^2}\right) = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2$$

We can call $\widetilde{\sigma^2}$ the **uncorrected sample variance**

Standard Error of Sample Mean

(Why divide by $n-1$?)

- Only $n-1$ independent pieces of information in $\{X_i - \overline{X}\}$ since $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$
- Given $\{X_1 - \overline{X}, \ldots, X_{i-1} - \overline{X}, X_{i+1} - \overline{X}, \ldots, X_n - \overline{X}\}$, you can calculate $X_i - \overline{X}$
- you used one "degree-of-freedom" when you used the data to calculate $\overline{X}$
- If $\overline{X}$ was obtained from a *different sample*, then you should divide by n, not $n-1$, to get an unbiased estimator for σ^2

Standard Error of Sample Mean

For our data, we have

```
x <- dat$earn
N <- length(x)
muhat <- mean(x)
s2hat <- var(x)
muhatse <- sqrt(s2hat/N)
cat("sample mean:", round(muhat,3), " s.e.:", round(muhatse,3), "\n")
```

```
sample mean: 29.232  s.e.: 0.368
```

Consistency

$$E(\overline{X}) = \mu \text{ and } Var(\overline{X}) = \frac{\sigma^2}{n} \to 0 \text{ as } n \to \infty$$

As $n \to \infty$, sample mean "converges" to μ

Convergence in Probability: A sequence of random variables Y_n, $n = 1, 2, ...,$ converges in probability to c if for any $\epsilon > 0$, we have

$$\lim_{n \to \infty} \Pr\left(|Y_n - c| \geq \epsilon \right) = 0 \, .$$

We say $Y_n \xrightarrow{p} c$

An estimator is **consistent** if it converges in probability to the true value of the parameter it is estimating

Consistency

Under our stated assumptions, the sample mean is consistent for the population mean

Khinchine's Weak Law of Large Numbers (WLLN) If $\{X_i\}_{i=1}^n$ is iid with $E(X_i) = \mu < \infty$ for all i, then

$$\overline{X}_n \xrightarrow{p} \mu$$

where $\overline{X}_n$ is the sample mean based on n observations.

- There are many "Laws of Large Numbers" each stating different conditions under which the sample mean is consistent
- "Weak" refers to the kind of probabilistic convergence used here (there are others)
- Bias and variance going to zero is actually "convergence in mean square", but this implies convergence in probability

Consistency (Simulation Example)

Suppose 200 people each took independent random samples of size n from population

Suppose population is well-represented by Chi-Sq(1) distribution (mean $= 1$)

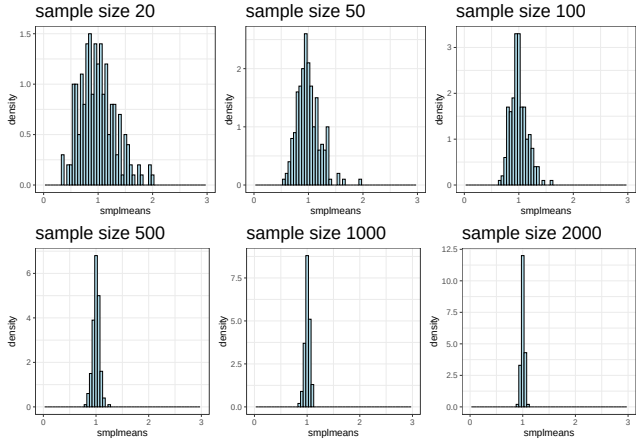
Plot distribution of the 200 sample means for sample sizes
$n = 20, 50, 100, 500, 1000, 2000$

- distribution of samples means if 200 people drew sample of size 20
- distribution of samples means if 200 people drew sample of size 50

etc.

Plots demonstrate unbiasedness and consistency

Consistency (Simulation Example)



Consistency

An important property of convergence in probability: if $g(.)$ is continuous, and $X_n \xrightarrow{p} c$, then $g(X_n) \to g(c)$

Suppose we want to estimate μ^2. Use $\widehat{\mu^2} = \overline{X}_n^2$?

- $\overline{X}_n^2$ is **not** an unbiased estimator of μ^2, since

$$Var(\overline{X}_n) = E(\overline{X}_n^2) - E(\overline{X}_n)^2 = E(\overline{X}_n^2) - \mu^2 \quad \Rightarrow \quad E(\overline{X}_n^2) = \mu^2 + Var(\overline{X}_n) > \mu^2$$

- $\overline{X}_n^2$ is a consistent estimator of μ^2, since

$$\overline{X}_n \xrightarrow{p} \mu \quad \Rightarrow \quad \overline{X}_n^2 \xrightarrow{p} \mu^2$$

Consistency

$\widetilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2$ is consistent for σ^2

Proof:

- X_i iid with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 \Rightarrow X_i^2$ iid with $E(X_i^2) = \sigma^2 + \mu^2$
- $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} \sigma^2 + \mu^2$ and $\overline{X}_n^2 \xrightarrow{p} \mu^2$
- Therefore $\widetilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2 \xrightarrow{p} \sigma^2 + \mu^2 - \mu^2 = \sigma^2$

$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ is also consistent for σ^2 since $\widehat{\sigma^2} = \underbrace{\frac{n}{n-1}}_{\to 1 \text{ as } n\to\infty} \widetilde{\sigma^2}$

Hypothesis Testing (Two-Sided)

Suppose we want to test

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_A : \mu \neq \mu_0$$

Intuitive Idea:

- If $\mu = \mu_0$ we expect $\hat{\mu}$ to be "near" μ_0
- If $\hat{\mu}$ is far from μ_0, perhaps $H_0 : \mu = \mu_0$ is incorrect
- If $\hat{\mu}$ is "too far" from μ_0, take this as statistical evidence that $\mu \neq \mu_0$

But how far is too far?

Hypothesis Testing (Two-Sided)

Assume for the moment that $X_i \overset{iid}{\sim} \text{Normal}(\mu_0, \sigma^2)$, $i = 1, \ldots, n$

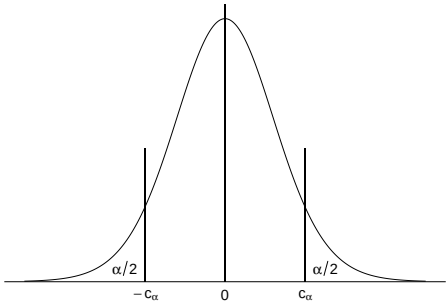
We have

$$X_i \overset{iid}{\sim} \text{Normal}(\mu_0, \sigma^2) \implies \overline{X}_n \sim \text{Normal}\left(\mu_0, \frac{\sigma^2}{n}\right)$$

$$\implies \frac{(\overline{X}_n - \mu_0)}{\sqrt{\sigma^2/n}} \sim \text{Normal}(0, 1)$$

$$\implies \underbrace{\frac{(\overline{X}_n - \mu_0)}{\sqrt{\widehat{\sigma^2}/n}}}_{t\text{-statistic}} \sim \text{t}(n-1)$$

Hypothesis Testing (Two-Sided)



Reject H_0 if $t > c_\alpha$ or $t < -c_\alpha$, where c_α is such that $\alpha = 0.01, 0.05, 0.10$ ("level of significance")

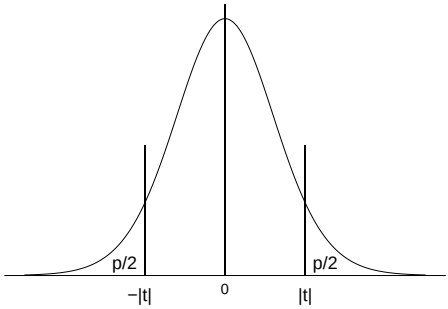
i.e., reject if $\Pr(|t| > c_\alpha) < \alpha$ given $\mu = \mu_0$ (Prob of rejecting correct null is α)

Hypothesis Testing (Two-Sided)

```
nVal <- c(20, 50, 100, 200, 400)
alphaVal <- c(0.01, 0.05, 0.1)
critVal <- matrix(rep(0,length(nVal)*length(alphaVal)), ncol = length(nVal))
colnames(critVal) <- paste0("n=",nVal)
rownames(critVal) <- paste0("alpha=",alphaVal)
for (i in 1:length(alphaVal)){
  for (j in 1:length(nVal)){
    critVal[i, j] = qt(1-alphaVal[i]/2, df=nVal[j]-1)
  }
}
round(critVal,3)
```

```
            n=20  n=50 n=100 n=200 n=400
alpha=0.01 2.861 2.680 2.626 2.601 2.588
alpha=0.05 2.093 2.010 1.984 1.972 1.966
alpha=0.1  1.729 1.677 1.660 1.653 1.649
```

Hypothesis Testing (Two-Sided)



Equivalently, reject $H_0 : \mu = \mu_0$ if "p-value" p in figure above is less than α, i.e., if the probability of obtaining a test statistic more extreme than the observed t-stat is less than α

Asymptotic Normality

When $n \to \infty$, the t-distribution converges to the Normal(0,1)

Then critical values $c_{0.01}$, $c_{0.05}$ and $c_{0.10}$ are 2.576, 1.96 and 1.645 respectively

- What if X_i doess not have normal distribution? Then t-statistic does not have t distribution.

However, we have the following result

Lindeberg-Levy Central Limit Theorem: If $\{X_i\}_{i=1}^n$ are iid with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$ for all i, then

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \text{Normal}(0, \sigma^2)$$

Asymptotic Normality (Simulation Example)

Note: cannot talk about the distribution of $\overline{X}_n$ as $n \to \infty$ since distribution collapses into single point

- Subtract μ to center it
- Multiply by $\sqrt{n}$ so that variance does not go to zero

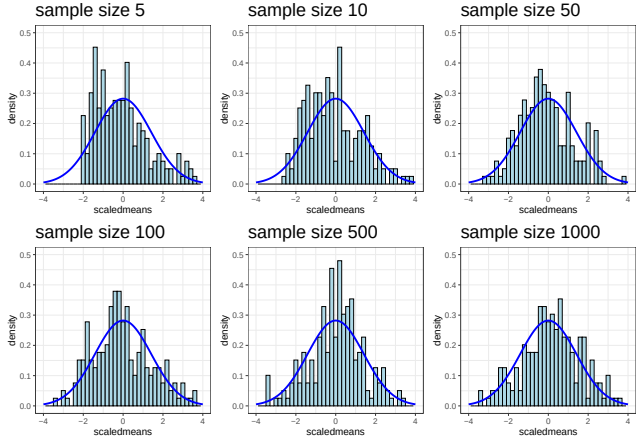
We can talk about distribution of $\sqrt{n}(\overline{X}_n - \mu)$ at $n \to \infty$

Continuation of simulation example of 200 people drawing independent samples from $\chi^2(1)$ population with sample sizes

$n = 5, 10, 50, 100, 500, 1000$

Plot distribution of $\sqrt{n}(\overline{X}_n - \mu)$ (here $\mu = 1$)

Asymptotic Normality (Simulation Example)



Hypothesis Testing (Two-Sided)

- "$\xrightarrow{d}$" means **convergence in distribution**
- when n is large, pdf of LHS is approximately the pdf of the Standard Normal
- Can also be shown that

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{\widehat{\sigma^2}}} = \frac{\overline{X}_n - \mu}{\sqrt{\widehat{\sigma^2}/n}} \xrightarrow{d} \text{Normal}(0,1)$$

  You can replace $\widehat{\sigma^2}$ with $\widetilde{\sigma^2}$ or any other consistent estimator of σ^2

That is, when n is large, can make the approximation $t \overset{a}{\sim} \text{Normal}(0,1)$, where $\overset{a}{\sim}$ means "approximately distributed", even when X_i is not normally distributed

Hypothesis Testing (Two-Sided) Example

Suppose we want to test

$H_0 : \mu = 30$ vs $H_A : \mu \neq 30$

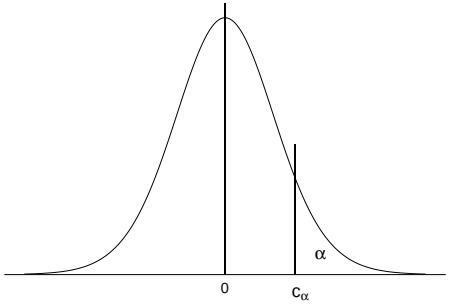
```
x <- dat$earn; n <- length(x); muhat <- mean(x); s2hat <- var(x)
tstat <- (muhat - 30)/sqrt(s2hat/n)
pval_t <- 2*pt(abs(tstat), df=n-1, lower.tail = FALSE)
pval_n <- 2*pnorm(abs(tstat), lower.tail = FALSE)
cat("t-stat:", tstat)
cat("\n p-value (t-dist):", pval_t)
cat("\n p-value (Standard Normal):", pval_n)
```

```
t-stat: -2.086885
 p-value (t-dist): 0.0369496
 p-value (Standard Normal): 0.03689851
```

Reject at 0.05 significance level but not 0.01

Hypothesis Testing (One-Sided)

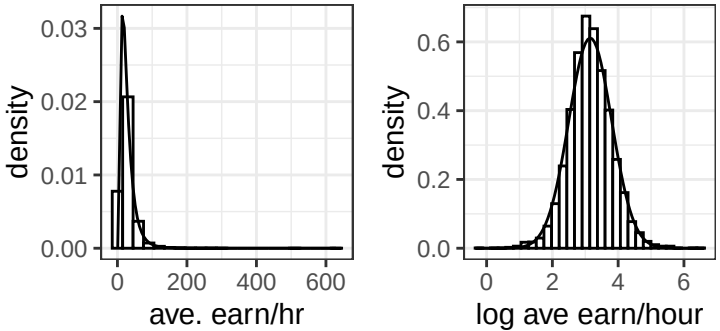
e.g., $H_0 : \mu \leq \mu_0$ vs $H_A : \mu > \mu_0$



Reject μ_0 if t-statistic is greater then c_α where c_α is that value such that $\Pr(t > c_\alpha) = \alpha$ under the null, $\alpha = 0.01, 0.05, 0.10$.

Estimation Again

Should we have worked with `log(earn)` instead of `earn`?



Estimation Again

Sample mean of $earn$ is unbiased and consistent for population mean (no problem in this regard)

Finite sample distribution of sample mean may not be normal or even symmetric in smaller samples because of high skew in distribution of X_i

- Concerns about hypothesis tests
- Harder to interpret standard errors

Better to work with $\ln earn$? To estimate $E(earn)$, probably not

When we get to regression, it will be easier and more sensible to work with $\ln earn$ instead of $earn$

Estimation Again

If you *do* work with $\ln earn$ with intention to convert results to $earn$, how do you do it?

To help us think about this, let's assume

$$\ln X_i \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2) \text{ for all } i$$

where X_i is earnings of individual i (seems reasonable!)

Then $X_i \overset{iid}{\sim}$ Log-normal(μ, σ^2) for all i

- $E(X_i) = e^{\mu+\frac{1}{2}\sigma^2} = e^{\mu}e^{\frac{1}{2}\sigma^2}$
- $Median(X_i) = e^{\mu}$

Estimation Again

Can estimate $\mu = E(\ln X)$ and $\sigma^2 = Var(\ln X)$ in the usual way

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \ln X_i \quad \text{and} \quad \widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n} (\ln X_i - \hat{\mu})^2$$

To convert mean and variance of $\ln X$ back to mean and variance of X

- estimate of mean hourly earnings: $e^{\widehat{\mu}} e^{\frac{1}{2}\widehat{\sigma^2}}$ (Not $e^{\widehat{\mu}}$)
- estimate of median hourly earnings: $e^{\widehat{\mu}}$

Also need to compute s.e. (use bootstrap?)

Estimation Again

Rough description of bootstrap idea

- Imagine again $R = 200$ people ($r = 1, \ldots, 200$) each collecting individual size n samples from population, and calculated sample means for their own samples $\overline{X}_{n,r}$
- You can calculate the standard error for the sample mean using

$$\frac{1}{R-1}\sum_{r=1}^{R}(\overline{X}_{n,r} - \overline{\overline{X}_n})^2$$

  where $\overline{\overline{X}_n}$ is the sample mean of the $R = 200$ sample means
- You can similarly get the standard error for the sample median

Estimation Again

Bootstrap idea: Create $R = 200$ "bootstrap" samples by re-sampling with replacement from your sample, i.e., from $\{X_1, \ldots, X_n\}$, get

$$\{X_1^{(1)}, \ldots, X_n^{(1)}\} \rightarrow \overline{X}_{n,1}, \; Med(X)_{n,1}$$
$$\vdots$$
$$\{X_1^{(r)}, \ldots, X_n^{(r)}\} \rightarrow \overline{X}_{n,r}, \; Med(X)_{n,r}$$
$$\vdots$$
$$\{X_1^{(R)}, \ldots, X_n^{(R)}\} \rightarrow \overline{X}_{n,R}, \; Med(X)_{n,R}$$

Calculate standard error of $\{\overline{X}_{n,r}\}_{r=1}^{R}$ and $\{Med(X)_{n,r}\}_{r=1}^{R}$ as standard error of your sample mean and sample median

Estimation Again

We apply this idea to get s.e. of mean and median of $earn$ after mean and variance of $\ln earn$

```
x <- log(dat$earn)
ln_mu_2_mu <- function(m, v){exp(m+0.5*v)}
ln_mu_2_md <- function(m, v){exp(m)}
m <- mean(x)
v <- var(x)
earnmean <- ln_mu_2_mu(m,v)
earnmed <-  ln_mu_2_md(m,v)
set.seed(456)
R <- 200                ## Bootstrap replication sample
bvars <- bmeans <- bmeds <- rep(NA, R)     ## To store the bootstrapped statistics
for (r in 1:R){
  xsmpb <- sample(x, 4946, replace=T)  # Sample with replacement from orig. smp.
  m1 <- mean(xsmpb)   # mean of bootstrap sample of ln(earn)
  v1 <- var(xsmpb)    # variance of boostrap sample of ln(earn)
  bmeans[r] <- ln_mu_2_mu(m1,v1)  # convert to mean of earn, and store
  bmeds[r]  <- ln_mu_2_md(m1,v1)   # convert to median of earn, and store
}
```

Estimation Again

```
cat("mean hr. earn.:   ", round(earnmean, 3),
    " s.e.:", round(sqrt(var(bmeans)),3), "\n")
cat("median hr. earn.: ", round(earnmed, 3),
    " s.e.:", round(sqrt(var(bmeds)),3),"\n")
```

```
mean hr. earn.:    28.907  s.e.: 0.305
median hr. earn.:  23.361  s.e.: 0.215
```

Earlier we calculated the mean of $earn$ directly and got

```
cat("mean hr. earn.:   ", round(muhat,3), " s.e.:", round(muhatse,3))
```

```
mean hr. earn.:    29.232  s.e.: 0.368
```