ECON207 Session 1

Math/Stats Review

Anthony Tay

Corrected Version: 20 Aug 2024

ECON207 Course Objectives

- Second course in UG econometrics
- Go deeper into theoretical foundations of OLS estimation of linear regression model
  - when it works well
  - when it doesn't work so well (or not at all)
  - how to use the models
  - using language of matrix algebra (needed for further work)
- Introduction to more advanced topics
  - instrumental variables
  - time series regressions
  - panel data
  - limited dependent variable models

Session 1

- A Bit of Math
  - Summation notation, probability prerequisites
  - We will cover more math throughout the course, as needed
- Statistics Review
  - Estimation
  - Hypothesis testing
- Course Administrative Arrangements
  - Course webpage vs Course eLearn page, Grading, Assignments

Session 1.1

Session 1.1 Math Review

- Summation Notation
- Probability Prerequisites

Summation Notation

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

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

Two Rules:

- $\sum_{i=1}^n (a_i + b_i) = \sum_{i=1}^n a_i + \sum_{i=1}^n b_i$
- $\sum_{i=1}^n c a_i = c \sum_{i=1}^n a_i$ 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)

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

```
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

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

Some Probability Prerequisites

Random variable, probability distribution function, mean (expected value) and variance, median

If X, Y are random variables, and a, b are constants

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

Some Probability Prerequisites

- X and Y independent: $f_{X,Y}(x,y) = f_X(x)f_Y(y)$
- X and Y independent $\Rightarrow Cov(X,Y) = 0$ but opposite implication need not hold
- Some distributions:
  - Normal (Gaussian) "Normal(μ, σ^2)"
  - Chi-sq "$\chi^2(v)$"
  - Student-t "$t(v)$"
  - Snedecor's F "$F(u,v)$"

If X and Y are Normal variables, then $aX + bY$ is Normal

More concepts/results to come...

Session 1.2

Session 1.2 Statistics Review

- Population vs Model vs Sample
- Evaluating Estimators
  - Unbiased Estimators
  - Efficiency
  - Consistency
- Estimator Standard Errors
- Hypothesis Testing

Statistics Review

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

e.g. Population of interest: Non-institutional employed civilians aged 16 and above in US in 2018

Population Characteristics of Interest:

1. "Representative" Hourly Earnings
2. Variation in Hourly Earnings across Population
3. Relationship between Hourly Earnings and Years of Schooling (Next week)

Random sample of n individuals from this population

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 make another 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

```r
library(tidyverse)
library(patchwork)
library(latex2exp)
dat <- read_csv("data\\earnings2019.csv", show_col_types=FALSE)
head(dat,3)
```

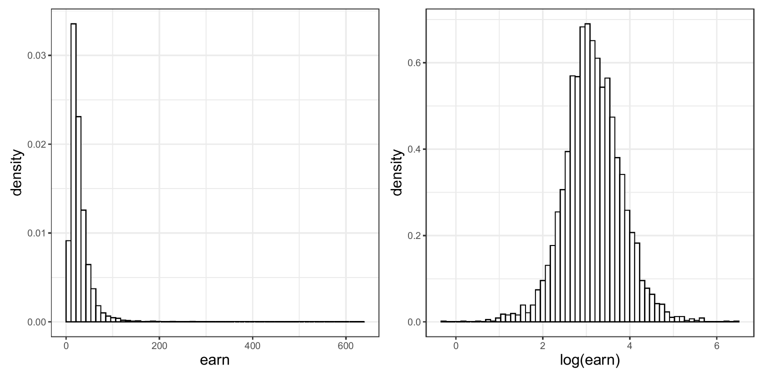
```
# A tibble: 3 x 11
    age height  educ feduc meduc tenure  wexp race   male  earn totalwork
  <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <chr> <dbl> <dbl>     <dbl>
1    59     67    12     3     3      5    30 White     0 36.3       1652
2    43     63    10     4     3      7    13 White    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


Data Example (Summary of Selected Variables)

```r
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)



Statistical Model

Statistical Model — A stylized description of the population and your sample $\{Y_i\}_{i=1}^n$

E.g. $Y_i \overset{iid}{\sim}$ Normal(μ, σ^2) where

- Y_i is earnings for individual i
- "iid" stands for independently and identically distributed (another interpretation of "random sample")

Not a good model!

Better for `log(earn)` than `earn`, but let's stick with `earn` for the moment

Statistical Model

It turns out we don't need to specify distribution fully

We can assume

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

Very general model! Assumes only that:

- sample is a random sample
- population is well-represented by *some* distribution with a mean and a variance (there are some distributions without finite mean / variance)

Suppose we want to estimate μ (population mean) and σ^2 (population variance)

Statistical Estimators

Since μ is $E(Y)$ and σ^2 is $Var(Y) = E((Y - E(Y))^2) = E(Y^2) - E(Y)^2$, suppose we decide

- $\hat{\mu} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ "Sample Mean"
- $\widetilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \overline{Y}^2$ (*we'll give this a name soon...*)

Is this a good idea?

We need to define what "good" means...

Bias

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

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

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

- 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.

Bias

Unfortunately, $\widetilde{\sigma^2}$ is a (downward) biased estimator of σ^2

Proof:

- Since $Var(Y_i) = E(Y_i^2) - E(Y_i)^2$, we have $E(Y_i^2) = \sigma^2 + \mu^2$
- Since $Var(\overline{Y}) = E(\overline{Y}^2) - E(\overline{Y})^2$, and $\overline{Y}$ is unbiased, we have $E(\overline{Y}^2) = Var(\overline{Y}) + \mu^2$

  Furthermore, we have $Var(\overline{Y}) = \dfrac{\sigma^2}{n}$:

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

  Therefore

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

Bias

Fortunately, in this case, there is an obvious unbiased estimator:

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

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

(Why divide by $n-1$?)

- Only $n-1$ independent pieces of information in $\{Y_i - \overline{Y}\}$ since $\sum_{i=1}^{n}(Y_i - \overline{Y}) = 0$
- Given $\{Y_1 - \overline{Y}, \dots, Y_{i-1} - \overline{Y}, Y_{i+1} - \overline{Y}, \dots, Y_n - \overline{Y}\}$, you can calculate $Y_i - \overline{Y}$
- you used one "degree-of-freedom" when you used the data to calculate $\overline{Y}$
- If $\overline{Y}$ was obtained from a *different sample*, then you should divide by n, not $n-1$, to get an unbiased estimator for σ^2

Estimator Standard Error

We should also try to get some idea of the size of estimation error:

We have already shown $Var(\overline{Y}) = \dfrac{\sigma^2}{n}$

Can replace σ^2 with its estimate: $\widehat{Var(\overline{Y})} = \dfrac{\widehat{\sigma^2}}{n}$

Standard error of sample mean: s.e.$(\overline{Y}) = \sqrt{\dfrac{\widehat{\sigma^2}}{n}}$

Estimator Standard Error

What is the "standard error for $\widehat{\sigma^2}$"?

Not conventionally computed as part of analysis

- Focus usually on the mean
- sample variance usually computed in order to compute standard error of the sample mean
- Nonetheless, a valid question
  - all estimates come with estimation error
  - good exercise!

Estimator Standard Error

Approach 1 (not a good one in this circumstance):

If we assume $Y_i \overset{iid}{\sim}$ Normal(μ, σ^2), then it can be shown that

$$\frac{(n-1)\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-1) \text{ which has a variance of } 2(n-1)$$

Then

$$Var\left(\widehat{\sigma^2}\right) = \frac{\sigma^4}{(n-1)^2}2(n-1) = \frac{2\sigma^4}{n-1}.$$

We can replace σ^2 with $\widehat{\sigma^2}$ to get

$$s.e.(\widehat{\sigma^2}) = \sqrt{\frac{2\left(\widehat{\sigma^2}\right)^2}{n-1}}$$

Estimator Standard Error

For our data, we have

```
y <- dat$earn; N <- length(y)
muhat <- mean(y); s2hat <- var(y)
muhatse <- sqrt(s2hat/N); s2hatse <- sqrt(2*s2hat^2/(N-1))
cat("sample mean:", round(muhat,3), " s.e.:", round(muhatse,3), "\n")
cat("sample variance:", round(s2hat,3),
    " s.e.:", round(s2hatse,3), "(don't trust this s.e.)\n")
```

```
sample mean: 29.232  s.e.: 0.368
sample variance: 670.651  s.e.: 13.487 (don't trust this s.e.)
```

The s.e. of the sample variance obtained here should not be trusted, since it is based on a formula derived assuming the data is Normally distributed, but our data is *far* from Normally distributed

Estimator Standard Error

Approach 2: hunker down and derive a formula for the variance of the sample variance *without assuming* Normality. *There is a formula* (we'll omit the proof :))

$$Var(\widehat{\sigma^2}) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right) \text{ where } \mu_4 = E((Y-E(Y))^4)$$

- μ_4 can be estimated by $\widehat{\mu_4} = (1/n)\sum_{i=1}^n (Y_i - \overline{Y})^4$
- If Y_i is normally distributed, then $\mu_4 = 3\sigma^4$ and $Var(\widehat{\sigma^2})$ reduces to $2\sigma^4/(n-1)$

```
mu4 = (1/N)*sum((y-mean(y))^4)
VV <- (1/N)*(mu4 - (N-3)/(N-1)*s2hat^2)
cat("sample variance:", round(s2hat,3), "    s.e. of sample variance:", round(sqrt(VV),3))
```

```
sample variance: 670.651     s.e. of sample variance: 95.358
```


Estimator Standard Error (The Bootstrap)

Approach 3: The Bootstrap

If R people obtained indp. random samples from pop. and calculated $\mu^{(r)}$ and $\widehat{\sigma^2}^{(r)}$

We can estimate standard error as s.e.$(\widehat{\sigma^2}) = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(\widehat{\sigma^2}^{(r)} - \overline{\widehat{\sigma^2}})^2}$

Idea of the bootstrap: *resample* from $\{Y_1, \dots, Y_n\}$ with replacement to get

$$\{Y_1^{(b)}, \dots, Y_n^{(b)}\} \text{ for } b = 1, \dots, B$$

Calculate for each bootstrap sample: $\widehat{\sigma^2}^{(b)}$ and then calculate

$$\text{bootstrap s.e.}(\widehat{\sigma^2}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}(\widehat{\sigma^2}^{(b)} - \overline{\widehat{\sigma^2}})^2}$$

Estimator Standard Error (The Bootstrap)

Can do the same for s.e. of the mean and the median!

```
set.seed(456)
B <- 200                ## Bootstrap replication sample
bmeans <- bvars <- bmeds <- rep(NA, B)     ## To store the bootstrapped vars, means, medians
for (b in 1:B){
  ysmpb <- sample(y, 4946, replace=T)  # Sample with replacement from orig. smp.
  bmeans[b] <- mean(ysmpb) # can do the same for the mean!
  bvars[b] <- var(ysmpb)   # bootstrapped sample variances
  bmeds[b] <- median(ysmpb) # can do the same for the medians!
}
cat("sample mean: ", round(muhat, 3), " s.e.:", round(muhatse,3),
    " bootstrap s.e.:", round(sqrt(var(bmeans)),3),"\n")
cat("sample var.: ", round(s2hat, 3), " s.e.:", round(s2hatse,3),
    " bootstrap s.e.:", round(sqrt(var(bvars)),3),"\n")
cat("sample median.: ", round(median(y), 3),  " bootstrap s.e.:", round(sqrt(var(bmeds)),3),"\n")
```

```
sample mean:  29.232  s.e.: 0.368  bootstrap s.e.: 0.357
sample var.:  670.651  s.e.: 13.487  bootstrap s.e.: 100.867
sample median.:  23  bootstrap s.e.: 0.314
```


Efficiency

Smaller estimator variance is better than larger estimator variance

Qn: Are there other unbiased estimators for μ with smaller variance?

(Partial answer, limiting ourselves to unbiased *linear* estimators)

Linear estimator for μ: estimator of the form $\tilde{\mu} = \sum_{i=1}^{n} w_i Y_i$

Unbiased of $\tilde{\mu}$ requires $\sum_{i=1}^{n} w_i = 1$

$$E(\tilde{\mu}) = E\left(\sum_{i=1}^{n} w_i Y_i\right) = \sum_{i=1}^{n} w_i E(Y_i) = \mu \sum_{i=1}^{n} w_i = \mu \text{ if } \sum_{i=1}^{n} w_i = 1$$

Efficiency

E.g.,

- sample mean is a linear unbiased estimator: weights $w_i = 1/n$, $i = 1, \dots, n$, sums to one.
- $\tilde{\mu}_1 = \frac{2}{n(n+1)} Y_1 + \cdots + \frac{2i}{n(n+1)} Y_i + \cdots + \frac{2n}{n(n+1)} Y_n = \sum_{i=1}^{n} \frac{2i}{n(n+1)} Y_i$

  $\tilde{\mu}_1$ is a linear estimator for μ, and unbiased since weights sum to one

$$\sum_{i=1}^{n} w_i = \sum_{i=1}^{n} \frac{2i}{n(n+1)} = \frac{2}{n(n+1)} \sum_{i=1}^{n} i = \frac{2}{n(n+1)} \frac{n(n+1)}{2} = 1 .$$

- $\tilde{\mu}_2 = y_n$ is a linear unbiased estimator

Efficiency

Under assumed conditions, **sample mean has smallest variance among all linear unbiased estimators** "Best Linear Unbiased"

Proof: Let $\tilde{\mu} = \sum_{i=1}^{n} w_i Y_i$ where $\sum_{i=1}^{n} w_i = 1$. Let $w_i = \frac{1}{n} + v_i$.

Since w_i sum to one, v_i sum to zero. Then

$$\begin{aligned} Var(\tilde{\mu}) &= \sum_{i=1}^{n} \left(\frac{1}{n} + v_i\right)^2 Var(Y_i) = \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n^2} + \frac{2v_i}{n} + v_i^2\right) \\ &= \frac{\sigma^2}{n} + \frac{2\sigma^2}{n} \sum_{i=1}^{n} v_i + \sigma^2 \sum_{i=1}^{n} v_i^2 = \frac{\sigma^2}{n} + \sigma^2 \sum_{i=1}^{n} v_i^2 \geq Var(\overline{Y}) . \end{aligned}$$

Equality holds only if $\sum_{i=1}^{n} v_i^2 = 0$, i.e., $v_i = 0$ for all $i = 1, \dots, n$, i.e., when $w_i = 1/n$

MSE and the Bias-Variance Tradeoff

Choosing BLU estimators places priority on unbiasedness

Alternative measure of quality of estimator — Mean Square Estimator Error

$$\begin{aligned} MSE(\hat{\theta}) &= E((\hat{\theta}-\theta)^2) \\ &= Var(\hat{\theta}-\theta) + (E(\hat{\theta}-\theta))^2 \\ &= Var(\hat{\theta}) + (E(\hat{\theta})-\theta)^2 \\ &= \text{Estimator Variance} + (\text{Estimator Bias})^2 \end{aligned}$$

Choosing estimator to minimize MSE allows for **bias-variance trade-off**

Can show that if $Y_i \overset{iid}{\sim}$ Normal(μ, σ^2), then $MSE(\widetilde{\sigma^2}) < MSE(\widehat{\sigma^2})$ (exercise)

Consistency

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

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

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

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

We say $X_n \xrightarrow{p} c$

An estimator is **consistent** if it converges in probability to what 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 $\{Y_i\}_{i=1}^{n}$ is iid with $E(Y_i) = \mu < \infty$ for all i, then

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

where $\overline{Y}_n$ 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

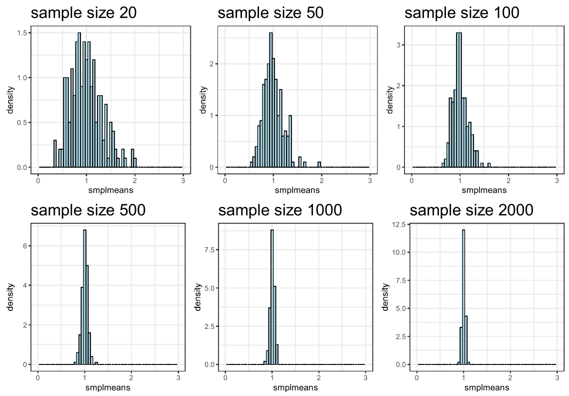
Suppose population is well-represented by Chi-Sq(1) distribution (mean $= 1$)

Plot distribution of sample mean for $n = 20, 50, 100, 500, 1000, 2000$

```
set.seed(1701)
Persons <- 200
MaxSampleSize <- 2000
AllSamples <- rchisq(Persons*MaxSampleSize, df=1) %>% matrix(ncol=Persons)
smplsizes <- c(20, 50, 100, 500, 1000, 2000)
plots1 <- vector("list", length=6)
for (i in 1:length(smplsizes)){
  n <- smplsizes[i]
  means <- colMeans(AllSamples[1:n,])
  datmeans <- data.frame(smplmeans=means)
  plots1[[i]] <- ggplot(data=datmeans, aes(x=smplmeans)) +
    geom_histogram(aes(y=..density..), color="black", fill="lightblue", binwidth=0.05) +
    labs(title = paste("sample size", smplsizes[i])) + xlim(0,3) +
    theme_bw() + theme(plot.title = element_text(size=20))
```

Consistency (Simulation Example)

```
(plots1[[1]] | plots1[[2]] | plots1[[3]]) / (plots1[[4]] | plots1[[5]] | plots1[[6]])
```



Consistency

Also, we say that $X_n \xrightarrow{p} Y_n$ if $X_n - Y_n \xrightarrow{p} 0$

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

- Suppose we want to estimate μ^2. A consistent estimator is $\hat{\mu}^2 = \overline{Y}^2$

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

Consistency

Note that $\overline{Y}^2$ is **not** an unbiased estimator of μ^2, since

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

Jensen's Inequality:

- If $g(.)$ is convex, then $E(g(X)) \geq g(E(X))$
- If $g(.)$ is concave, then $E(g(X)) \leq g(E(X))$
- Equality holds if $g(.)$ is linear

e.g. $g(x) = x^2$ is strictly convex

Consistency

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

Proof:

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

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

Hypothesis Testing (Two-Sided)

Suppose we want to test

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

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 $Y_i \overset{iid}{\sim} \text{Normal}(\mu_0, \sigma^2)$, $i = 1, \ldots, n$

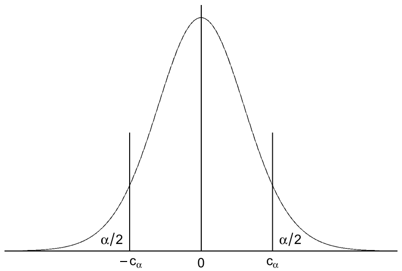
We have

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

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

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

Hypothesis Testing (Two-Sided)



Reject H_0 if $t > c_\alpha$ or $t < -c_\alpha$, where c_α is such that $\alpha = 0.01, 0.05, 0.10$

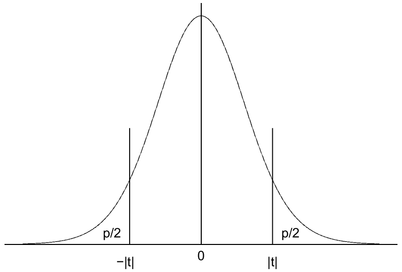
i.e., reject if $\Pr(|t| > c_\alpha) < \alpha$ given $\mu = \mu_0$ (Prob of rejecting correct null)

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" $\Pr(|t| > c_\alpha)$ 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 Y_i is not Normally distributed? Then t-statistic does not have t distribution.

However, we have the following result

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

$$\sqrt{N}(\overline{Y} - \mu) \xrightarrow{d} \text{Normal}(0, \sigma^2)$$

Asymptotic Normality (Simulation Example)

Continuation of Simulation Example (200 people drawing independent samples from population)

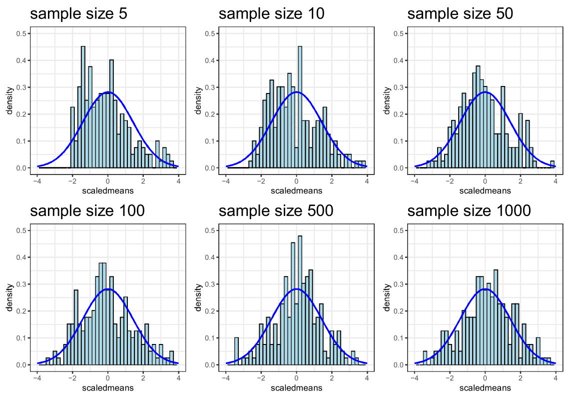
$n = 5, 10, 50, 100, 500, 1000$

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

```
plots2 <- vector("list", length=6)
smplsizes <- c(5, 10, 50, 100, 500, 1000)
for (i in 1:length(smplsizes)){
  n <- smplsizes[i]
  means <- colMeans(AllSamples[1:n,])
  datmeans <- data.frame(scaledmeans=sqrt(n)*(means-1))
  plots2[[i]] <- ggplot(data=datmeans, aes(x=scaledmeans)) +
    geom_histogram(aes(y=..density..), color="black", fill="lightblue", binwidth=0.2) +
    stat_function(fun=dnorm, args = with(dat, c(mean=0, sd=sqrt(2))), color="blue", size=1) +
    xlim(-4, 4) +  ylim(0, 0.5) + labs(title = paste("sample size", smplsizes[i])) +
    theme_bw() + theme(plot.title = element_text(size=20))
}
```


Asymptotic Normality (Simulation Example)

```
(plots2[[1]] | plots2[[2]] | plots2[[3]]) / (plots2[[4]] | plots2[[5]] | plots2[[6]])
```



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

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

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

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

Hypothesis Testing (Two-Sided) Example

For our data

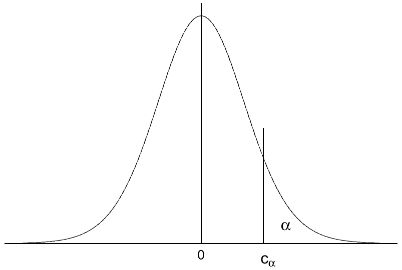
$H_0 : \mu = 30$ vs $H_A : \mu \neq 30$

```
y <- dat$earn; N<- length(y); muhat <- mean(y); s2hat <- var(y)
t <- (muhat - 30)/sqrt(s2hat/N)
pval_t <- 2*pt(abs(t), df=N-1, lower.tail = FALSE)
pval_n <- 2*pnorm(abs(t), lower.tail = FALSE)
cat("t-stat:", t)
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
```


Hypothesis Testing (One-Sided)

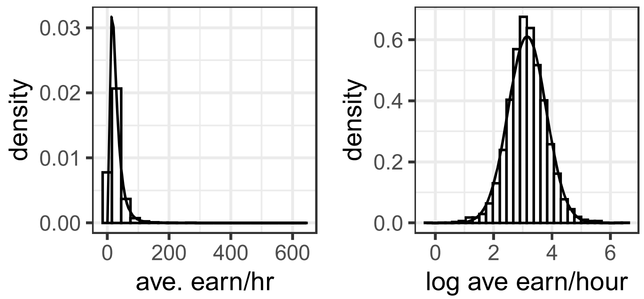
$H_0 : \mu < \mu_0$ vs $H_A : \mu \geq \mu_0$



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

To help us think about this, let's assume

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

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

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

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

Estimation Again

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

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

But we are interested in mean and variance of Y, not $\ln Y$ — must convert back!

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

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

Estimation Again

```
y <- log(dat$earn)
n2ln_mu <- function(m, v){exp(m+0.5*v)}
n2ln_vr <- function(m, v){exp(2*m + v)*(exp(v) - 1)}
n2ln_md <- function(m, v){exp(m)}
m <- mean(y)
v <- var(y)
earnmean <- n2ln_mu(m,v)
earnvar <-  n2ln_vr(m,v)
earnmed <-  n2ln_md(m,v)
set.seed(456)
B <- 200                  ## Bootstrap replication sample
bvars <- bmeans <- bmeds <- rep(NA, B)      ## To store the bootstrapped statistics
for (b in 1:B){
  ysmpb <- sample(y, 4946, replace=T)  # Sample with replacement from orig. smp.
  m1 <- mean(ysmpb)   # mean of bootstrap sample of ln(earn)
  v1 <- var(ysmpb)    # variance of boostrap sample of ln(earn)
  bmeans[b] <- n2ln_mu(m1,v1)  # convert to mean of earn, and store
  bvars[b] <- n2ln_vr(m1,v1)   # convert to variance of earn, and store
  bmeds[b] <- n2ln_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("var. hr. earn.:  ", round(earnvar, 3),
    " s.e.:", round(sqrt(var(bvars)), 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
var. hr. earn.:   443.902  s.e.: 20.8
median hr. earn.:  23.361  s.e.: 0.215
```

Session 1.3

Session 1.3 Course Arrangements

- Course Arrangements
  - Webpages, reading material, software
  - Grading system

Course Arrangements

- Course webpage vs course eLearn page
- Course Notes
- Software: R
  - Not covered in class (learn by playing with code supplied)
  - Needed for Assignment
  - NOT EXAMINABLE (no stress!)

Course Arrangements (Evaluation)

- **Individual Assignments 50%**
  - Short Weekly Review Questions (20%), graded based on submission, feedback via detailed answer sheet
  - Three longer assignments (30%), graded in detail.
- **Exam 40%**
  - Closed book, calculators allowed, **no cheat sheet**
- **Class and Forum Participation 10%**
  - ask/answer questions in class
  - ask/answer questions on forum page
  - post typos and errors on forum page

Roadmap

- **This Session 1: Statistics Review**
- *Next Session 2: Simple Linear Regression*
- Session 3: Estimator Standard Errors; Multiple Linear Regression
- Session 4: Matrix Algebra
- Session 5: OLS using Matrix Algebra
- Session 6: Hypothesis Testing
- Session 7: Prediction
- Session 8: Instrumental Variable Regression
- Session 9: Logistic and Other Regressions
- Session 10: Panel Data Regressions
- Session 11: Introduction to Time Series
- Session 12: Time Series Regressions