36-463/663: Multilevel & Hierarchical Models

R as a Statistical Calculator

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Outline

- Announcements & Office Hours
- What is Statistics For?
- Distributions as Models
- Confidence Intervals
- Hypothesis Tests
- G&H Ch's 3-4 start reading now
  - I will not cover everything in the chapters
  - You will need to read & try some things on your own!

Announcements & Office Hours

- HW02: Due next Tue Sep 13, on Blackboard.
- Nick's Regular Office Hours (BH 132M):
  - Mon 5-6
- Brian's Regular Office Hours (132E Baker):
  - Tue, Thu 3-4
  - Some Tuesdays after class

What is Statistics For?

- Statistical inference is used to learn from incomplete or imperfect data.
  - **<u>Sampling model</u>**: the data are *<u>incomplete</u>* because of sampling.
    - E.g. estimate the opinions of the entire United States based on a sample of 1,000 respondents.
    - No random error in people's answers
    - *<u>Uncertainty</u>* arises from which & how many persons we ask
  - **<u>Measurement error model</u>**: the data are *<u>imperfect</u>* because of errors in measurement
    - Your test score is not an exact measure of your knowledge
    - In $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, the "linear part" is not an exact relationship between y and x... ϵ_i is the error.
    - *<u>Uncertainty</u>* arises because the pairs (x_i, y_i) have extraneous information in them, for estimating β_0 and β_1.

What is Statistics For?

- There can be measurement error in survey sampling. It is ideally dealt with by careful design and pre-testing of questions, to make it go away.
  - Sometimes measurement error models needed anyway (NAEP)
- There can be sampling in measurement error problems
  - In the London schools example, looks like not all students from each of the 38 schools were in the data set.
  - Most regression models combine measurement and sampling uncertainty:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
ϵ_i is drawn from a $N(0, \sigma^2)$ distribution

Distributions as Models

- Statistical distributions are tools for modeling uncertainty
  - Distributions can represent the population from which we are sampling and/or the method by which we sample (*sampling models*)
  - Distributions can represent the messy or unknown part of the process of generating data (*generative models*)
- Statistical distributions can be used either for sampling or generative models
- It's important to know some common situations that different distributions are good at modeling!

Distributions as Models: Normal

- Arises when an observation is a sum of many similar small independent contributions:

If Z is a sum of independent contributions

$$Z = Z_1 + Z_2 + \cdots Z_n = \sum_{i=1}^{n} Z_i$$

then, approximately, $Z \sim N(\mu_z, \sigma_z^2)$, with

$$\begin{aligned} \mu_z &= E[Z] &= \sum_{i=1}^{n} E[Z_i] &= \sum_{i=1}^{n} \mu_{z_i} \\ \sigma_z^2 &= \mathrm{Var}(Z) &= \sum_{i=1}^{n} \mathrm{Var}(Z_i) &= \sum_{i=1}^{n} \sigma_{z_i}^2 \end{aligned}$$

as long as each $\sigma_{z_i}^2$ is small relative to σ_z^2, and the μ_{z_i}'s aren't too different from each other.

Aside: Building Sums in R…

```
> old.opt <- options(digits=3)

> (x <- rnorm(10))
 [1] -0.632 -0.207 -1.745  0.150 -1.098 -0.528  0.236 -0.518 -0.377 -0.364
> (xsum <- sum(x))
[1] -5.08
```

- We want to produce many (perhaps 100's) of sums like this
  - For example, to draw a histogram of sums of 10 x's
- Doing by hand and storing each one is tedious
- How can we automate this process?
  - Produce a vector of sums…

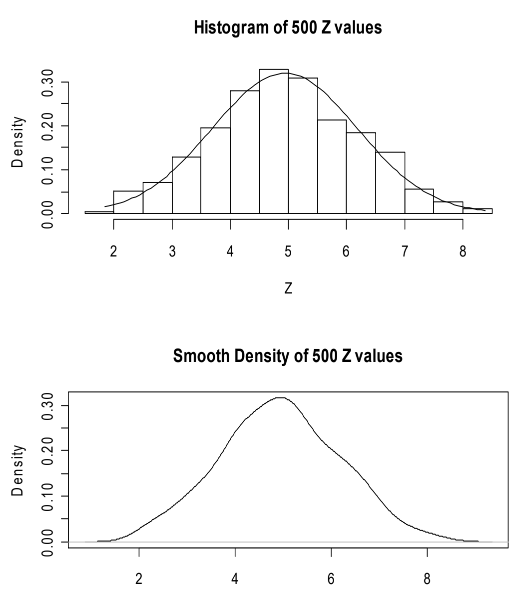
Aside: Building Sums in R…

```
> old.opt <- options(digits=3)

> (x <- rnorm(10))
 [1] -0.632 -0.207 -1.745  0.150 -1.098 -0.528  0.236 -0.518 -0.377 -0.364
> (xsum <- sum(x))
[1] -5.08

> (xdata <- matrix(rnorm(5*10),ncol=10))
       [,1]  [,2]   [,3]   [,4]   [,5]   [,6]   [,7]    [,8]   [,9]   [,10]
[1,] -2.889 0.621 -1.129 -2.116  0.720  0.770  0.780 -0.4469  0.180  1.0100
[2,] -0.826 0.131  1.605  0.885 -0.388  1.133 -1.086  0.1395 -1.443 -0.6977
[3,]  0.741 0.917  1.119  0.906 -1.058 -0.192  0.788 -0.0322  0.196 -0.0779
[4,]  0.278 1.853 -0.286  0.100  1.162  0.192 -1.314  1.0876 -1.839  0.2338
[5,]  0.068 0.925  0.420  0.152 -1.015 -1.605  1.908  0.9469  1.040 -0.5053
> (xsum <- apply(xdata,1,sum))
[1] -2.499 -0.547  3.306  1.466  2.335

> options(old.opt)
```


Distributions as Models: Normal

If $z_i \sim \text{Unif}(-\frac{1}{4}, \frac{3}{4})$
Then $z = \sum_{i=1}^{n} z_i$ is approximately normal

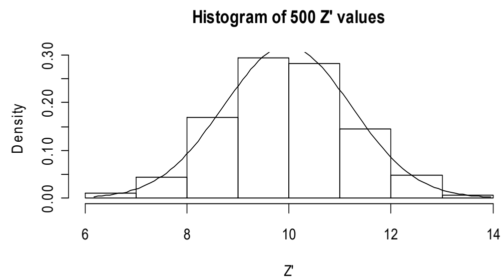
```
Z_i <- matrix(runif(n=500*20,
  min=-0.25,max=0.75),ncol=20)
Z <- apply(Z_i,1,sum)

par(mfrow=c(2,1))
hist(Z,probability=T)
x <- seq(min(Z),max(Z),length=100)
lines(x, dnorm(x,mean(Z),sd(Z)))
plot(density(Z), main=
  "Smooth Density of Z",xlab="Z")
```

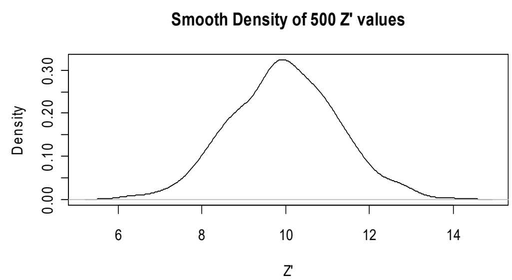


Distributions as Models: Normal

If $z'_i \sim \text{Unif}(0,1)$
Then $z' = \sum_{i=1}^{n} z'_i$ is approximately normal

```
ZZ_i <- matrix(runif(n=500*20,
  min=0,max=1),ncol=20)
ZZ <- apply(ZZ_i,1,sum)

par(mfrow=c(2,1))
hist(ZZ,probability=T,
  main="Histogram of Z‘ ",xlab="Z‘ ")
x <- seq(min(ZZ),max(ZZ),length=100)
lines(x, dnorm(x,mean(ZZ),sd(ZZ)))
plot(density(ZZ), main=
  "Smooth Density of Z‘ ", xlab="Z‘ ")
```





Distributions as Models: Normal

If the means are not similar this will not work:

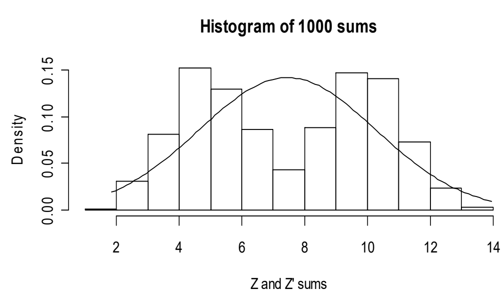
$$z_i \sim \text{Unif}(-\frac{1}{4}, \frac{3}{4}),\ E[z_i] = \frac{1}{4}$$

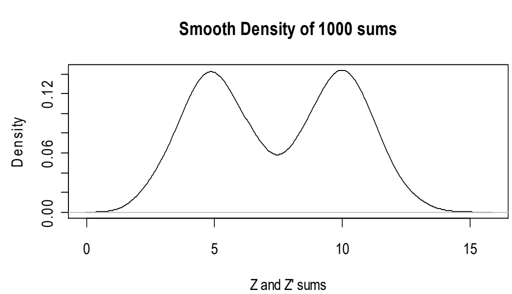
$$z'_i \sim \text{Unif}(0,1),\ E[z'_i] = \frac{1}{2}$$

Combining 500 samples of each of

$$z = \sum_{i=1}^{n} z_i, \text{ and } z' = \sum_{i=1}^{n} z'_i$$

will not produce a normal distribution.





Aside: the d, p, q and r functions…

- **dnorm(x,mean,sd)** produces values of the (norm)al (d)ensity
- **pnorm(x,mean,sd)** prodices values of the (norm)al (p)robability P[Z $\leq$ x] (i.e. the normal cdf)
- **qnorm(p,mean,sd)** produces the (q)uantile x for which P[z $\leq$ x] = p (i.e., the inverse normal cdf)
- **rnorm(n,mean,sd)** produces n independent (r)andom draws of Z
- *Every distribution that R knows about has a **d,p,q** and **r** function!*

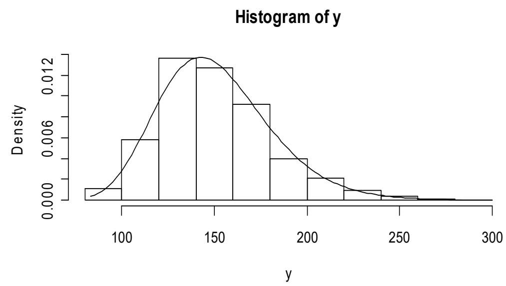
Distributions as Models: Log-Normal

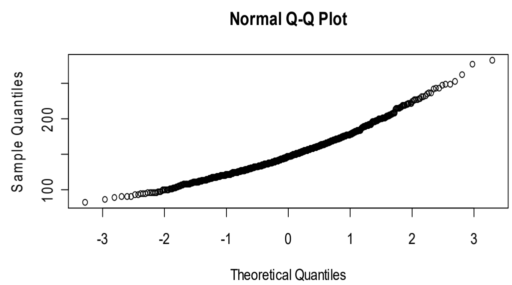
- Some distributions (dollars earned, distance ball thrown, etc.) are naturally skewed right.
- A common “remedy” is to take the logarithm of the data.
  - We will always use the natural log (log base e, where e=2.71828… is Euler’s constant)
  - Since there will never be any confusion, we will just write log(x)
    - Not ln(x), not $\log_e(x)$
- This “fix” leads to the “log-normal” distribution

Distributions as Models: Log-Normal

- Y $\sim$ lognormal(μ,σ^2) iff log(Y) $\sim$ normal(μ,σ^2)
- With a little calculus, can show that the density of Y is dnorm(log(y))/y

```
mean <- 5
sd <- 0.2
z <- rnorm(1000,mean,sd)
y <- exp(z) # so that log(x) ~ N(0,1)

par(mfrow=c(2,1))
hist(y,probability=T)
x <- seq(min(y),max(y),length=100)
lines(x,dnorm(log(x),mean,sd)/x)
qqnorm(y)
```

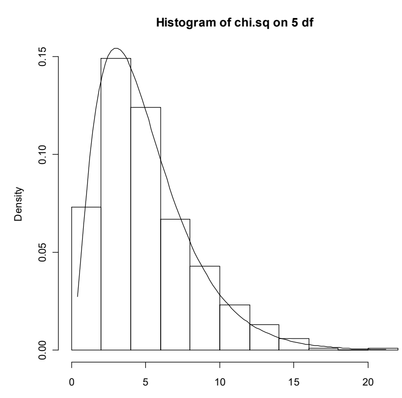




Distributions as Models: Chi-squared

- Chi-squared on k df is the sum of k N(0,1)^2's
  - Distribution of sample variance is a constant times a chi-squared
- Chi-squared also arises in
  - Likelihood ratio tests
  - Testing independence in tables of counts

```
df = 5
chi <- matrix(rnorm(500*df),ncol=df)
chi.sq <- apply(chi^2,1,sum)

hist(chi.sq,probability=T,main=
  paste("Histogram of chi.sq on",df,"df"))
plot(function(x) {dchisq(x,df=df)},add=T,
  from=min(chi.sq),to=max(chi.sq))
```



Distributions as Models: others…

- Binomial
- Beta
- Poisson
- Student's t
- Multivariate Normal
- Multinomial
- Dirichlet
- Gamma
- Wishart
- …and many more…

We do not have to memorize these for now, but don't be surprised when they arise!

Confidence Intervals: Normal Data

- A 100(1-α)% CI for the mean of a normal population based on a sample of size n is:

  (xbar + qt(α/2,n-1)·SE, xbar + qt(1-α/2,n-1)·SE)

```
> y <- c(35,34,38,35,37)
> n <- length(y)
> x.bar <- mean(y)
> se <- sd(y)/sqrt(n)
> (int.50 <- x.bar + qt(c(.25,.75),n-1)*se)
[1] 35.2557 36.3443
> (int.95 <- x.bar + qt(c(.025,.975),n-1)*se)
[1] 33.75974 37.84026
```

Conf. Intervals: Binomial Proportion

- A 100(1-α)% CI for a binomial proportion, based on a sample of size n is:

  (p.hat + qnorm(α/2)·SE, p.hat + qnorm(1-α/2)·SE)

```
> y <- 700
> n <- 1000
> p.hat <- y/n
> se <- sqrt (p.hat*(1-p.hat)/n)
> (int.95 <- p.hat + qnorm(c(.025,.975))*se)
[1] 0.6715974 0.7284026
> (int.95.approx <- p.hat + c(-2,2)*se)
[1] 0.6710172 0.7289828
```


Confidence Intervals: Simulation

- Suppose we survey men and women's attitudes toward death penalty
  - 375 of 500 men favor death penalty (75%)
  - 325 of 500 women favor death penalty (65%)
- The *ratio* of support of men to women is 0.75/0.65 = 1.15.
- How could we build a *confidence interval* for this ratio (as a way of estimating the ratio in the full population that these men and women were samped from)?

Confidence Intervals: Simulation

```
> n.men <- 500
> p.hat.men <- 0.75
> se.men <- sqrt (p.hat.men*(1-p.hat.men)/n.men)

> n.women <- 500
> p.hat.women <- 0.65
> se.women <- sqrt (p.hat.women*(1-p.hat.women)/n.women)

> n.sims <- 10000
> p.men <- rnorm (n.sims, p.hat.men, se.men)
> p.women <- rnorm (n.sims, p.hat.women, se.women)
> (ratio <- p.men/p.women)
    [1] 1.1363384 1.1389650 1.0696729 ... ... ...
 [9997] 1.1661268 1.1745934 1.0499191 1.1391952
> (int.95 <- quantile (ratio, c(.025,.975)))
    2.5%    97.5%
1.062888 1.251581
```


Hypothesis Testing

- Deciding about a ***null Hypothesis*** $\underline{H_0}$ vs an ***alternative Hypothesis*** $\underline{H_A}$
- Key question: is the data very unlikely under the null hypothesis?
  - If the data is unlikely under the null hypothesis this is evidence to reject H_0
  - If the data seem pretty likely under the null hypothesis, then we can't reject H_0
- Logic of tradit. hypothesis testing never allows us to accept H_0 or H_A, only to assess evidence against H_0, reject H_0 with some confidence

Hypothesis Testing by Eyeballing Confidence Intervals

- **If the parameter value under H_0 is not in the 95% confidence interval, we reject H_0 at level α=0.05.**
- *In the normal-data CI example*, let's test H_0: μ=35. Since the 95% CI was (33.8, 37.8), and 35 is in this interval, we fail to reject at α=0.05.
- *In the binomial proportion example*, let's test H_0: p=0.65. Since the 95% CI was (0.67, 0.73), we reject H_0 at α=0.05
- *In the death penalty example*, is it plausible that men and women think equally of the death penalty (H_0: ratio=1)? The 95% CI was (1.06, 1.25), so we would reject ratio=1 at the α=0.05 level.

Hypothesis Testing Using a Null Distribution

- A sample of 50 people are asked their favorite color and also asked to take an introversion / extroversion test.
- H_0: these are independent factors; H_A: dependent

Observed Counts	Blue	Red	Yellow	TOTAL
Introverted	5	20	5	30
Extroverted	10	5	5	20
TOTAL	15	25	10	50

Hypothesis Testing Using a Null Distribution

- The "expected" counts under H_0: independence are
  (row total)*(column total)/(grand total)

Expected Counts	Blue	Red	Yellow	TOTAL
Introverted	9	15	6	30
Extroverted	6	10	4	20
TOTAL	15	25	10	50

Hypothesis Testing Using a Null Distribution

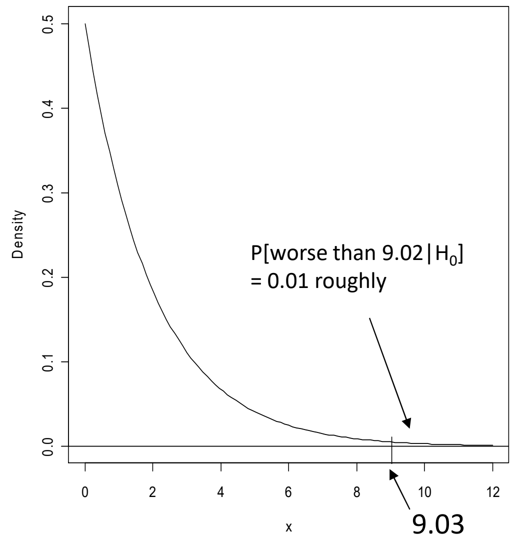
- The Chi-squared test statistic is

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{(obs_{ij} - exp_{ij})^2}{exp_{ij}} = 9.03$$

- The model for this statistic under H_0 is chi-squared on k df, where k=(rows-1)*(cols-1)=2
- Our data would be unlikely if our test statistic is far out in the tail of this model
  - If our data is unlikely under the H_0 model, reject H_0

Hypothesis Testing Using a Null Distribution

```
psy.table <- matrix(c(5,20,5,10,5,5),byrow=T,nrow=2)

row.sums <- apply(psy.table,1,sum)
col.sums <- apply(psy.table,2,sum)
expected <- outer(row.sums,col.sums)/sum(psy.table)

(chi.sq <- sum((psy.table-expected)^2/expected))
[1] 9.027778
plot(function(x) { dchisq(x,df=2) }, from=0, to=12,
  ylab="Density")
lines(c(chi.sq,chi.sq),c(-1,2*dchisq(chi.sq,df=2)))
abline(h=0)
pchisq(9.02,2,lower=F)
[1] 0.01099846
```



- Since P[worse data than 9.02 | H_0] = 0.01 (small), we can reject H_0: color preference and introversion aren't independent!

Hypothesis Testing Using Simulation

- We want to know if we can compare mean test scores of students in two schools. If the variances of the test scores in the two schoosl are similar, we can compare means with a two-sample t-test.
  - School A: n_A = 130 students, s^2_A = 25.1
  - School B: n_B = 120 students, s^2_B = 20.9
- There is an exact F-test (assuming the test scores are normally distributed) but rather than look it up, let's proceed by simulation.
  - H_0: $\sigma^2_A / \sigma^2_b = 1$
  - H_A: not

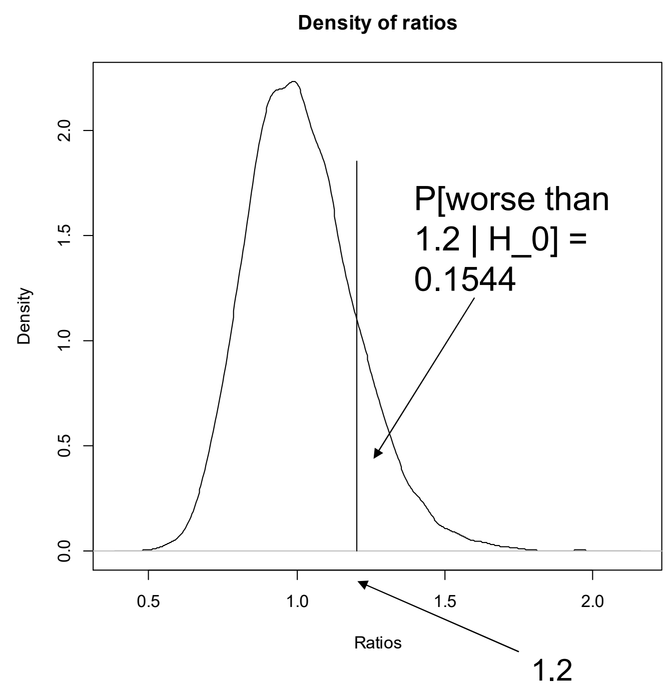
Hypothesis Testing Using Simulation

```
> nsims <- 10000
>
> n.A <- 130
> s2.A <- 25.1
>
> n.B <- 120
> s2.B <- 20.9
>
> (obsd.ratio <- 25.1/20.9)
[1] 1.200957
> s2.pooled <- (s2.A*(n.A-1) + s2.B*(n.B-1))/(n.A + n.B - 1)
> sims.A <- matrix(rnorm(nsims*n.A,0,sqrt(s2.pooled)),byrow=T,nrow=nsims)
> vars.A <- apply(sims.A,1,var)
>
> sims.B <- matrix(rnorm(nsims*n.B,0,sqrt(s2.pooled)),byrow=T,nrow=nsims)
> vars.B <- apply(sims.B,1,var)
>
> ratios <- vars.A/vars.B
```

```
> worse <- (obsd.ratio <= ratios)
>
> (pval <- sum(worse)/nsims)
[1]  0.1544
>
> plot(density(ratios),main="Density of ratios",xlab="Ratios",ylab="Density")
> lines(c(obsd.ratio,obsd.ratio),c(0,1.85))
> # see plot on next page…
```


Hypothesis Testing Using Simulation

- Since P[worse data than 1.2 | H_0] = 0.1544 (big), we cannot reject H_0: the variances in the two groups effectively the same.



Summary

- What is Statistics For?
- Distributions as Models
- Confidence Intervals
- Hypothesis Tests
- G&H Ch's 3-4 – start reading now
  - I will not cover everything in the chapters
  - You will need to read & try some things on your own!
- Office Hours
- HW02 Due next Tues Sep 13