CHAPTER 2: MATERNAL SMOKING AND INFANT DEATH

math 189 : Data Analysis and Inference : winter 2018

Jelena Bradic
http://www.jelenabradic.net
Assistant Professor, Department of Mathematics, University of California, San Diego
jbradic@ucsd.edu

Introduction

Data

Background

Investigations

Theory

Jelena Bradic http://www.jelenabradic.net Assistant Professor, Department of Mathematics, University of California, San Diego jbradic@ucsd.edu

Exploratory Data Analysis

Before making inferences from the data, it is essential to examine all your variables/data

Why?

To listen to the data:

* to catch mistakes
* to see patterns in the data
* to find violations of statistical assumptions
* to generate hypothesis
* ..... and because if you don't, you will have trouble later

Infant Death Tied to Premature Births

Introduction

Wednesday, March 5, 1995 // New York Times Low Weights not solely to blame

Surgeon General warning

"Smooking by pregnant women may result in fetal injury, premature birth and low birth weight"

* Objective: evaluate the surgeon general warning by comparing the birth weights of babies born to smokers and to non-smokers.
* Primary Data : part of the Child Health and Development Studies (CHDS) - data is collected of all pregnancies that occurred between 1960 and 1967 among women in Kaiser Health Plan of the San Francisco region.
* This data is known for its unexpected finding that ounce for ounce babies of smokers did not have a higher death rate than the babies of nonsmokers.
* Despite the surgeon general working: 15% of pregnant women smoke during pregnancy (study done in 1996).

Details of studies

Insurance providers

Information from "previous knowledge"

Exact information

* Epidemiological studies indicate that smoking is responsible for 150g of reduction in birth weight.
* Epidemiological studies indicate that smoking mothers are twice as likely as nonsmoking mothers to have a low-birth-weight baby (under 2500g).
* Babies maturity is measured by :

Definition

  ** birth weight
  ** baby's gestational age (0 – 40 weeks)

  Babies both early and born small have lower survival rates.

Jelena Bradic http://www.jelenabradic.net Assistant Professor, Department of Mathematics, University of California, San Diego jbradic@ucsd.edu

Data

Part of ReadME. file

* Data collected for our study is enlarged portion of the mentioned CHDS data.
* The data consists of all pregnancies that occurred between 1960 and 1967 among women in Kaised Health Plan in Oakland, California.
* The women in the study are all those that
  ** were enrolled in Kaiser Heath Plan ← enough money
  ** has obtained prenatal care in San Francisco area ← regularly see doctors, taken care
  ** and delivered in any of the Kaiser hospitals in Northen California ← environment was healthy.

limiting confounding factors

* At birth measurements are obtained: length, weight and head circumference.
* Our study is comprised of 1236 babies
  ** all the same gender: boys
  ** single births: no twins
  ** all lived at least 28 days ← Effect of smoking

other reason: die early ⇒ poor health
⇒ SIDS (just stop breathing)

↑ end of mature after there

Data matrix

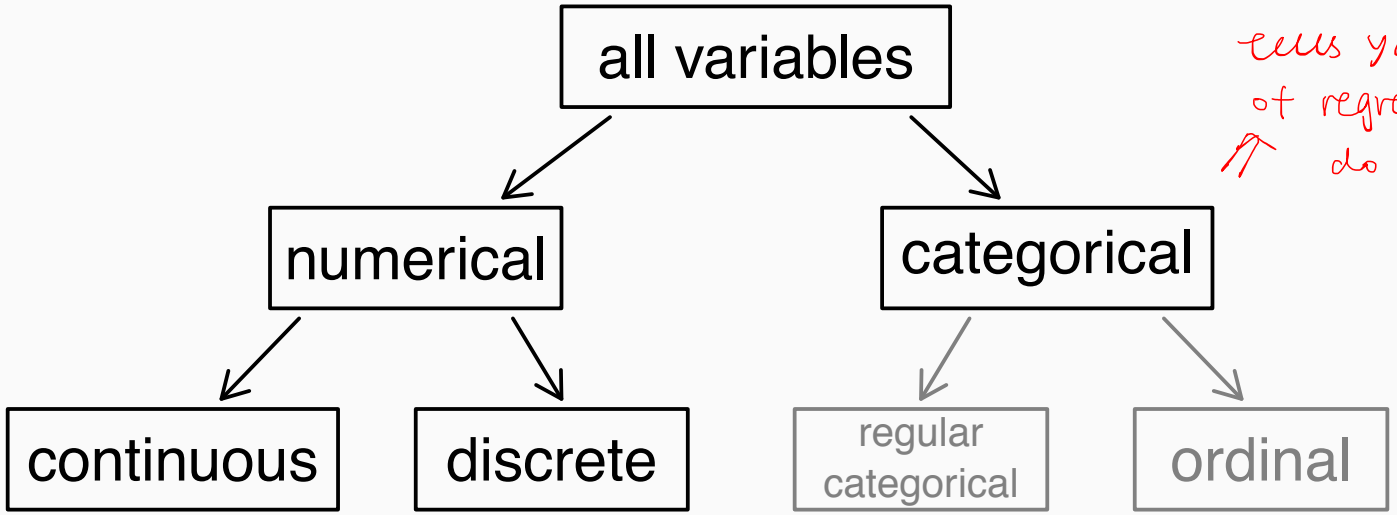
Information available to us is: birth weight and wether or not the mother smoked

variable ↓

Stu.	babies weight in ounces	smoking status
1	120	not-smoked
2	113	not-smoked
3	128	smoked
4	123	not-smoked
⋮	⋮	⋮
1236	117	not-smoked

← observation

Types of variables



tells you which type of regressions you should do

Least square

(Poisson, Binomial)

Generalized linear regression

Types of variables (cont.)

variable ↓

Stu.	babies weight in ounces	smoking status
1	120	not-smoked
2	113	not-smoked
3	128	smoked
4	123	not-smoked
⋮	⋮	⋮
1236	117	not-smoked

← observation

* smooking status:

Types of variables (cont.)

variable ↓

Stu.	babies weight in ounces	smoking status
1	120	not-smoked
2	113	not-smoked
3	128	smoked
4	123	not-smoked
⋮	⋮	⋮
1236	117	not-smoked

← observation

* `smooking status`: categorical

Practice

What type of variable is a `babies weight in ounces`?

* numerical, continuous
* numerical, discrete
* categorical, ordinal
* categorical

How to do that?

How did other people do that?

Jelena Bradic, http://www.jelenabradic.net, Assistant Professor, Department of Mathematics, University of California, San Diego, jbradic@ucsd.edu

What can we learn about the underlying science

Try doing this; ⇒ prepared for final

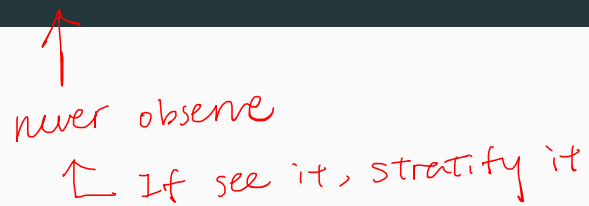
Research question: How does the fetal grow?

Facts: (later will need to be done / researched by ourselves.

* The typical gestational period for a baby is 40 weeks.
* Preterm delivery: baby born before 37 weeks.
* Some babies stay in utero up to 42 weeks.
* At 28 weeks, the fetus weights about 4 to 5 pounds and is about 40cm long.
* At 32 weeks, the fetus weights about 5 to 5.5 pounds and is about 45cm long.
* In the final weeks, baby gains 0.2 pounds a week.
* Most newborns range from 45 to 55cm in length and from 5.5 to 8.8 pounds.
* Babies born at term that weight less than 5.5 pounds are considered small for their gestational age.

Journal Article

What is all the fuss about Confounders



Confounders: Possible items that can mask the effect of smoking on the death rate.

Need to look at the history/literature of the problem itself: Rubella

* Before 1940's it was believed that there is no disease that mother has that can hurt the baby.
* Dr Norman Gregg, 1941, observed an "unusual" number of infants with congenital cataracts: all the mothers has contracted rubella in the first or second month of their pregnancy.
* He went to a statistician or two to confirm how "unusual " is really unusual.

More to the science

* It is commonly thought that carbon monoxide in cigarette smoke reduces the oxygen supplied to the fetus .
* The physiological effects of a decreased oxygen supply on fetal development are not completely understood.
* Steady supply of the oxygen is critical for the developing baby.
* It is hypothesized that to compensate for the decreased supply of oxygen, the placenta increases in surface area and in the number of blood vessels. This is believed to lead to “abruptia placenta” where the placenta brakes away from the uterine wall and results in a preterm delivery and fetal death.

Is the difference important

* If a difference is found between the birth weights of babies born to smokers and this born to nonsmokers, the Q is: is this important to the health and development of the baby (aka the death rate)?
* Four different death rates - fetal (die before born), neonatal (first 28 days after birth), perinatal (combined fetal and neonatal) and infant
  * Previous studies exists and show that although low birth weights are associated with an increase in the numbers of death shortly after birth, the babies of smokers tended to have much lower death rates than the babies of nonsmokers.
    ** Rates were not adjusted for possible confounders: mother's age for example ....

heaven effect
in stead of
weight effect

other than the shift in the mean
= distribution of weight in two groups
⇒ other than mean, survival rate is the same
heaven is not affected

Wilcox AJ, Skjaerven R. Birth weight and perinatal mortality: the effect of gestational age. American Journal of Public Health 1992;82(3):378-382.

two distributions have the same shape ⇒ no difference in the headed
↳ look at var, shape, mode

Check Assumptions

Objective Questions
↓
histograms → understand problems
↓
Understand if assumption holds

Is the difference important (cont.)

* In order to compare mortality rates of smokers and non-smokers, Wilcox and authors, advocate grouping babies by their gestational age, or by their relative birth weight (= birthweight - average birth weight of its group).
    * ../figures/Wilcox.pdf
    ** Because babies born to smokers tend to be smaller, the mortality curve is shifted to the right relative to the nonsmokers' curve.
    ** If the babies born to smokers are smaller but otherwise as healthy as babies born to nonsmokers, then the two curves in standard units should coincide.
    ** Wilcox and Russell found that the mortality rate of the smokers is higher.

Introduction

Data

Background

Investigations

Theory

hypothesis test
understand graph / method you are gonna use
use google scholar to search

Jelena Bradic http://www.jelenabradic.net Assistant Professor, Department of Mathematics, University of California, San Diego jbradic@ucsd.edu

Investigations

Big question:

What is the difference in weight between babies born to mothers who smoked during pregnancy and those who did not? Is this difference important to the health of the baby?

Instructions

- Summarize numerically the two distributions of birth weight for babies born to women who smoked during their pregnancy and for babies born to women who did not smoke during their pregnancy.
- Use graphical methods to compare the two distributions of birth weight.
- Compare the frequency, or incidence, of low-birth-weight babies for the two groups. How reliable do you think your estimates are? That is, how would the incidence of low birth weight change if a few more or fewer babies were classified as low birth weight?
- Asses the importance of the difference you found in your three types of comparisons (numerical, graphical, incidence). Summarize your findings and relate them to other studies.

qqplot, histogram

# of times something happened

Add noise to it? Don't change too much to check

How robust is the estimator?

How reliable the estimator is?

Discussion

↳ refleets back

Simulate

↳ Illustrate whether this is a good estimator.

Note: If you make separate plots for smokers and nonsmokers, be sure to scale the axes identically for both graphs.

only use words

Introduction

Data

Background

Investigations

Theory

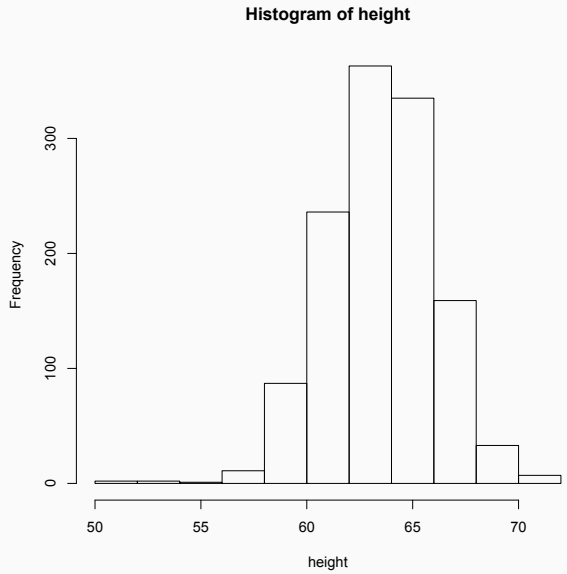
Jelena Bradic http://www.jelenabradic.net Assistant Professor, Department of Mathematics, University of California, San Diego jbradic@ucsd.edu

Summary Statistics

* Why: When analyzing a set of data, simple summaries of the list of numbers cab bring insight about the data.
* Example: The mean and the standard deviation are frequently used as the numerical summaries of the location and the spread of the data.
* Graphical summary: Histograms, box plots, and quantile-quantile plots often provide information about the distribution of the data (symmetry, modality, tails of the data).
* Data in lecture: We illustrate some of the summary statistics on the data from the 1236 families in the Child Health and Development Study (CHDS). Missing data is ignored.

Histogram

Figure below displays the heights of mothers in the CHDS study.



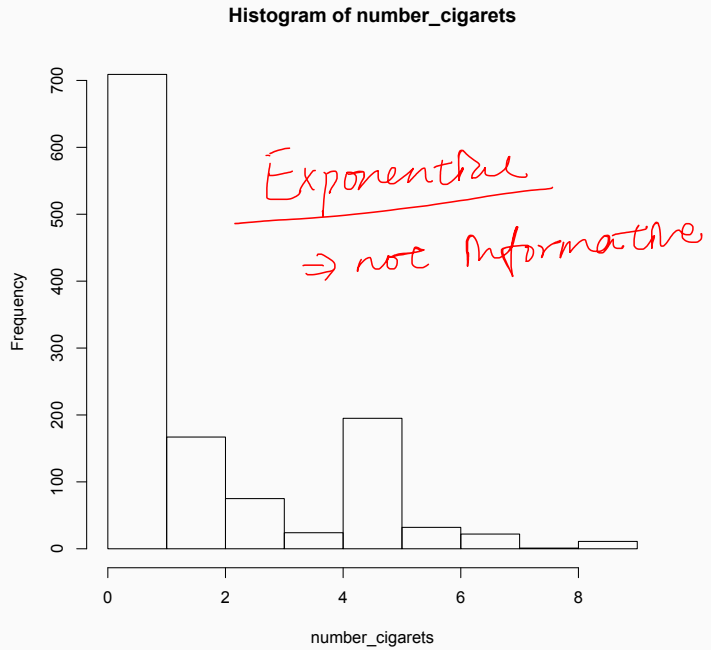
Histogram is unimodal and symmetric.
The distribution has one mode about 64 inches
The shape of the histogram to the left of the peak looks roughly like the miser image of the part of the histogram to the right of the peak.
Histograms indicate a few outliers (short mothers in the study).

Histogram (cont.)

Figure below displays the histogram for the number of cigarettes smoked per day of mothers who smoked during pregnancy in the CHDS study.

Histogram has two modes: around 5-10 and 20-30.

The distribution is asymmetric: right skewed



Histogram (cont.)

* Histograms can be used to answer distributional questions.

What proportion of the babies weight under 100 ounces or what percentage of the babies weight more than 138 ounces?

*

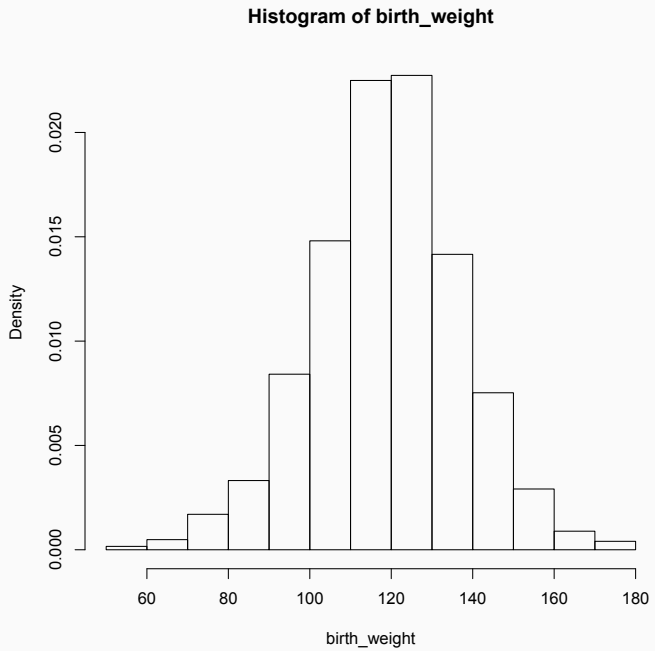
* From the histogram

$$\sum(\text{areas to the left of } 100)$$

14%

Note: 138 does not belong to the interval endpoints, hence we need to split the bar between 130-140 into 10 ounce subintervals: the whole bar contains 14.2%, so we estimate that each one-ounce subinterval contains roughly 1.4% of the babies

Figure below displays the histogram for the babies weight in the CHDS study.



Numerical Summaries

* Average: represents the center of the data distribution (balance point of the histogram) ; computed to give a ”center” around which the measurements in the data are distributed.
* Standard Deviation: how far individual value may vary from the center of the distribution.
* Lower Quartile: such number that at least 25% of the data falls bellow it and at least 75% of the data fall above it (when more than one value meets the criterion we take on the average of the two).

to the left

to the left

Normal Curve

Theory

The standard normal curve, known as the bell curve, sometimes provides a useful method for summarizing data.

* Standard normal curve is unimodal and symmetric around zero.
* It follows 68-95-99.7 rule:
  ** 68% of the area under the curve is within 1 unit of its center
  ** 95% of the area under the curve is within 2 units of its center
  ** 99.7% of the area under the curve is within 3 units of its center
* The areas are determined from the following analytical expression of the curve

density function

$$\frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\}.$$

* Traditionally,

cdf function of standard normal

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\}.$$

pnorm

numerical interpretation

* Many distributions of the data are approximately normal, and the 68-95-99.7 rule can be used as the informal check of normality.

Behold the power of the Central Limit Theorem

Theory

Theorem

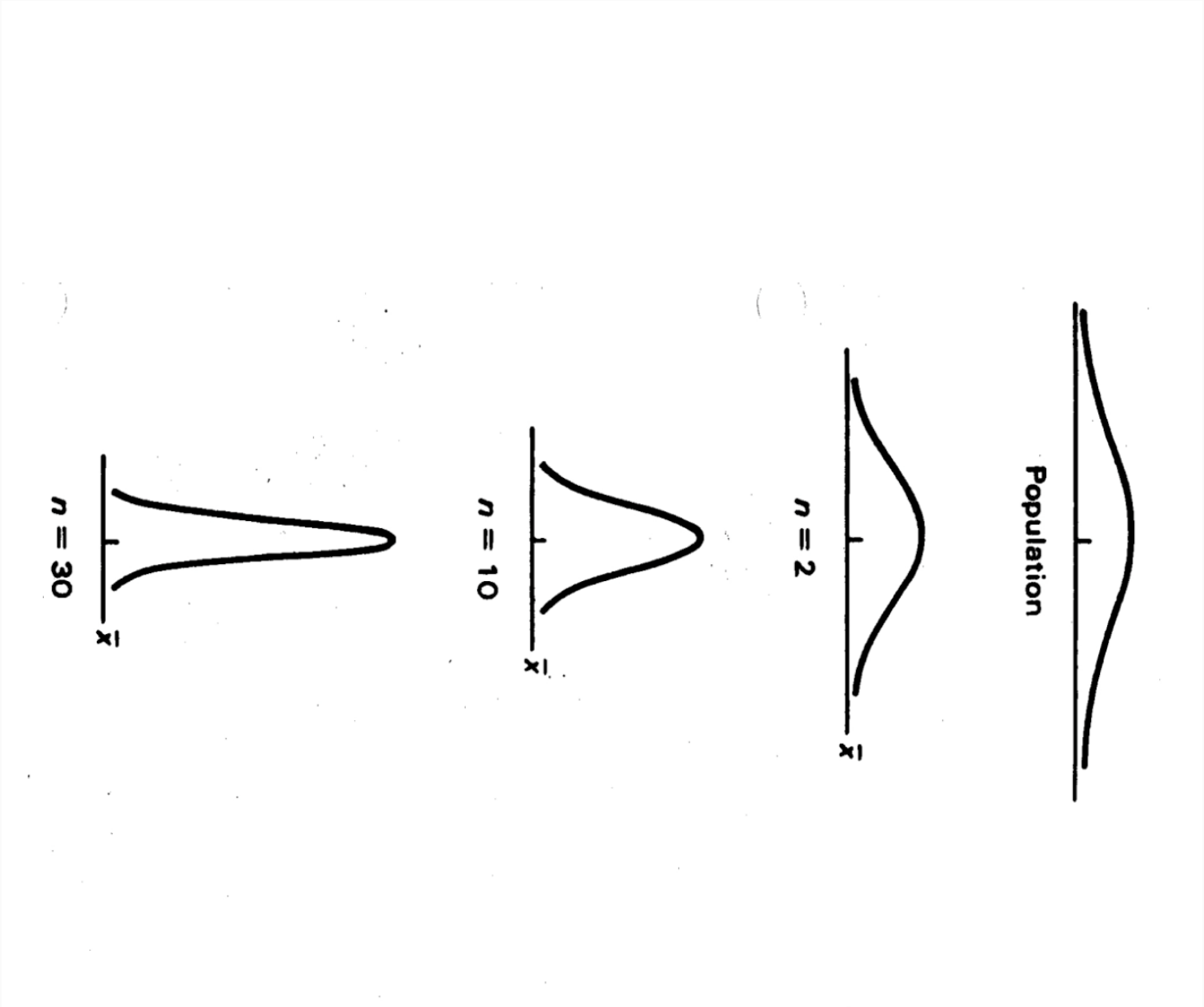
Let $X_1, X_2, \ldots, X_n$ be an i.i.d. random sample from a distribution with mean μ and standard deviation σ. If n is sufficiently large:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$$

$\bar{X}$ is closer and closer to μ

↳ independent, finite mean, var

From the figure we see that the shape of the population doesn't affect the shape of the sample average



Normal Curve (cont.)

Regression

try to standardize data

* If the histogram of the data looks normal, then the 68-95-99.7 rule should apply to properly standardized data
* To standardize the data, we subtract the mean from each observation and divide by standard deviation

$$\frac{X_i - \bar{X}}{S}$$

⇒ same numerical range

⇒ don't want data to be sensitive

with

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

($\bar{X}$: μ; S^2: σ^2)

** For birth weight we find that 69% of the babies have weights within 1 standard deviation of the average
** For birth weight we find that 96% of the babies have weights within 2 standard deviations of the average
** For birth weight we find that 99.4% of the babies have weights within 3 standard deviations of the average

Checks for normality

Goodness of Fit
whether two has same distribution

guess distribution

Simulations
Monte Carlo → compute integral / partial differential equations
Bootstrap

* More formal checks of normality are based on skewness and kurtosis.
* Skewness coefficient: the average of the third power of the standardized data

purpose
→ check for normality.

$$\text{skewness} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s}\right)^3$$

symmetry

applies for all distributions

* Kurtosis coefficient: the average of the fourth power of the standardized data

$$\text{kurtosis} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s}\right)^4$$

→ data that are standardized

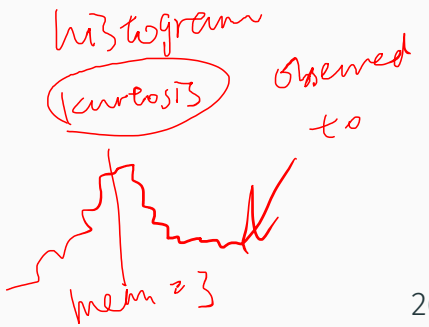
** For symmetric distributions, the skewness coefficient is 0.
** Kurtosis coefficient: how pronounced is the peak of the distribution.
** For normal distribution, the kurtosis coefficient is 3.

normal
→ kurtosis
= 3

skewed? close to
↳ skewness → 0?

Big or small departures of normality

* To decide wether a departure from normal distribution is big or small, simulation studies can be used.
* Simulation study: generates pseudo random numbers from a given distribution
* Check the similarity of the simulated data with the observed data.
* This may be used to show that a particular distribution would be unlikely to give us the data we see.
  ** Kurtosis of birth weight for the 484 babies born to smokers is 2.9.
  ** To check for normality repeat as follows 1000 times:
    1. Generate 484 pseudo-random observations from a normal distribution and calculate the kurtosis coefficient
    2. The following figure is a histogram of those kurtosis over 1000 repetitive computations and data generations as the above.



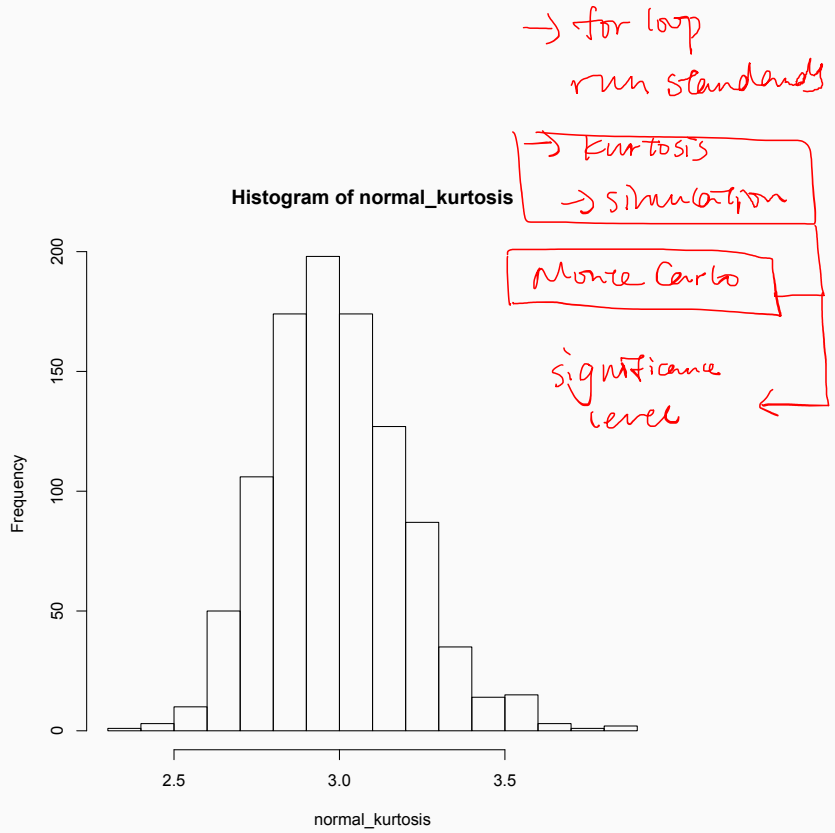
Big or small departures of normality

moment

```
normal_kurtosis=NULL
for(i in 1:1000)

normal_kurtosis[i]=kurtosis(rnorm(484))
hist(normal_kurtosis)
```

compute quantile of 2.9



From the figure we see that 2.9 is very typical kurtosis value of the normal distribution with 484 samples

Quantiles

Let X be a random variable. Then, p-th quantile of the random varaible X is defined as any number q satisfying

$$P(X \leq q) \geq p$$

and

$$P(X \geq q) \geq (1 - p)$$

Sometimes, q is denoted with $F^{-1}(p)$.

Sample quantiles are based on Order Statistics

Quantile-quantile plots

testing for tails

For standard normal distribution, the q-th quantile is z_q, where

$$\Phi(z_q) = q, 0 < q < 1.$$

The median, lower and upper quantiles are examples of quantiles. They are, respectively, 0.5,0.25 and 0.75 quantiles.

For the data $X_1, \ldots, X_n$, the sample quantiles are found by ordering the dat from smallest to largest. We denote this ordering by

$$X_{(1)}, \ldots, X_{(n)}.$$

Then $X_{(k)}$ is considered as the $k/(n+1)$-th sample quantile.

Note: we divide by $n + 1$ to keep q smaller than 1.

Comparing two samples

Suppose we have two samples of size n: $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$.

* If they were samples from the sample distribution, then the order statistics $X_{(1)}, \ldots, X_{(n)}$ and $Y_{(1)}, \ldots, Y_{(n)}$ would be estimates of the sample quantiles.
* Thus we expect $X_{(1)} \sim Y_{(1)}, \ldots, X_{(n)} \sim Y_{(n)}$

Quantile-quantile plots (cont.)

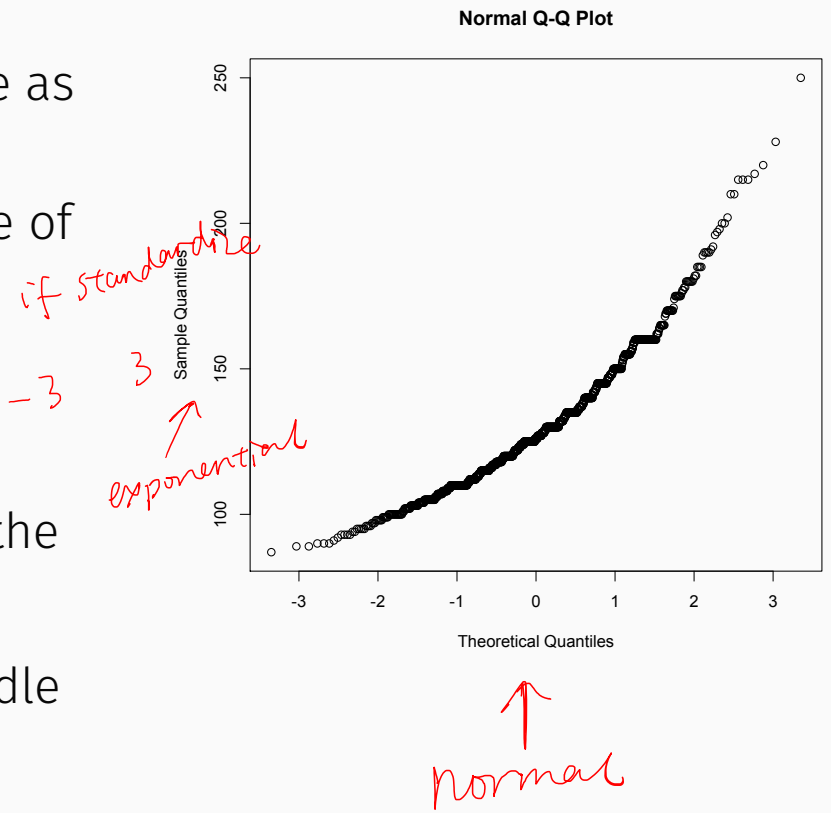
* Normal quantile plot provides a graphical means of comparing the data distributions to the normal.
* It plots pairs

$$(Z_{k/(n+1)}, X_{(k)}).$$

* If the plotted points point roughly on the line, then it indicates that the data have an approximate normal distribution.

Quantile-quantile plots (cont.)

* Departures of normality are indicated by systematic departures from a straight line.
  ** The upward curve in the plot identifies a long right tail, in comparison to normal.
* If the histogram of the data does not decrease as quickly in the right tail as the normal, this is indicated by an upward curve on the right side of the normal-quantile plot.
* Similarly, a long left tail is indicated by a downward curve to the left.
* Granularity of the data appears as stripes on the plot.
* Bimodality of the data appears as curved middle of the plot.



Quantile-quantile plots for comparing data distributions

* Quantile-quantile plots can be used for any distribution.
* Quantile-quantile plots for comparing data distributions compare two sets of data to each other by pairing their respective sample quantiles.
* Again, departure from a straight line indicates a difference in the shapes of the two distributions..
* When the distributions are identical, the plot should have intercept 0 and slope 1.
* If the two distributions are the same shape but have different means or standard deviation, then the plot should also be roughly linear but slope and intercept will not be 1 and 0.
  ** A nonzero intercept indicates a shift in the distributions.
  * Nonunit slope indicates a scale change.

Some recommendations

The aim of good data graphics

* Display data accurately and clearly.

Some rules for displaying data badly:

* Display as little information as possible
* Obscure what you do show (with chart junk)
* Make a pie chart (preferably in color and 3d)
* Use pseudo 3d and color gratuitously
* Use a poorly chosen scale