Principles of Inferential Statistics in Medicine (EPIB 607)

Erica E. M. Moodie

Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

erica.moodie@mcgill.ca
www.medicine.mcgill.ca/epidemiology/moodie/InferentialStats.htm

July 24, 2007

Part 1: Descriptive statistics: Data collection, description, and display

Course content

1. Descriptive statistics: Data collection, description, and display

- Types of data
- Visual summaries: histograms, stem & leaf plots, boxplots
- Numerical summaries: means, medians, variance
- Rescaling

Why perform descriptive statistics?

- Identify errors in measurement or data collection
  - Univariate: too high or low
  - Multivariate: strange combinations
- Characterize materials and methods
  - Describe subjects used in a study
- Summarize missing data
  - Univariate: how much is missing in each variable?
  - Multivariate: do any variables predict missingness?
- Assess the validity of assumptions needed for analysis
  - Distribution of the data
  - For multivariate: relationships between variables
  - For multivariate: explore the possibility of confounding

Why perform descriptive statistics?

- Hypothesis generation
  - Explore unexpected effects
  - Explore effects in different subgroups (e.g., is an effect similar in men and in women?)

Descriptive statistics

In this part of the course, we will focus on summarizing each variable separately.

Some general methods:

- What are the sampling methods?
  - Source of data? Location, time, selection (inclusion/exclusion) criteria
- What is the scientific meaning of each variable?
  - Demographic, exposure (treatment), clinical outcome, disease severity, ...?
- Compute summary measures of the distributions
  - Where appropriate, consider tables, means, medians, plots, etc.
  - The type of summary that is most appropriate depends on the data type of the variable

Types of data

Qualitative

- Non-numerical
- Binary (all or none, yes/no)
- Multi-category (ordered or unordered)
- Summarize with tables, proportions

Quantitative

- Numerical (measured)
- Discrete
- Continuous
- Summarize with histograms, measures of location and spread

Summarizing qualitative data

Binary variables:

- Only two possible values (e.g., yes/no, male/female, low birth-weight/normal birth-weight)
- It's usually easiest to assign numbers to these (e.g., yes = 1/no = 0, Male = 1 if the person is male and Male = 0 if the person is female)
- We can interpret averages of binary data, and differences and means of these ratios are scientifically meaningful quantities

Summarizing qualitative data

Categorical data:

- A finite number of possible values denoting categories (e.g., *occupation* is clerical/labourer/stay-at-home parent/professional/retired; *marital status* is single, married, co-habiting, divorced, separated, widowed)
- Categorical data can be sub-divided in **ordered** and **unordered** data.
- If data are ordered, the ordering may be partial (e.g., Pap smears results may fall into one of six ordered categories or by "indeterminate")
- Means, differences, and so on aren't well-defined for categorical variables

Quantitative data

Quantitative or numerical data:

- May be discrete (e.g., number of pregnancies, number of infections)
- May be continuous (e.g., weight)
- Note that we can summarize these data with means and ratios, but that interpretation of ratios may be scale specific (e.g., “twice as hot” is different in Celsius compared to Fahrenheit)

Quantitative data

Censored data:

- A special type of missing information is due to **censoring**
- **Right censoring**: We know only that the true value is greater than some threshold (e.g., time of remission: a person has lived cancer-free for 8 months when the study ends, so we know time of remission is longer than 8 months)
- **Left censoring**: We know only that the true value is less than some threshold (e.g., HIV RNA is below the limit of detection of a particular type of assay)
- **Interval censoring**: We know only that the true value occurs in some interval (e.g., date of infection: we know that a patient was uninfected at visit j and infected at visit $j + 1$, but we don't know when in that period he became infected)

Quantitative data

The bottom line?

You better think (think)
think about what you're trying to do to me.

– Aretha Franklin, *Think*

Displaying Numerical Data

Erica E. M. Moodie

We will begin with an example in birds. It is hypothesized that symmetry in animals (including people!) is favourable.

A study was conducted to see whether maternal stress affects the symmetry of her off-spring. Quail eggs randomly allocated to be injected with oil (a *control* or *placebo* condition) or with a steroid that is naturally secreted by female quails when stressed.

Once the quail eggs hatched, right and left legs were measured to compare symmetry.

Displaying Numerical Data

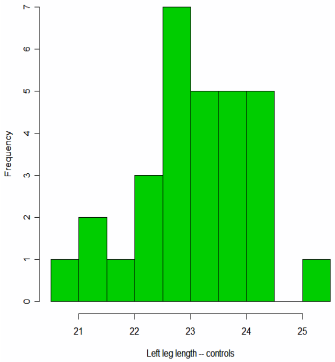
Control		Corticosterone	
Left leg	Right leg	Left leg	Right leg
23.96	24.24	24.12	24.4
23.38	23.76	25.04	25.86
24.08	23.66	22.34	22.66
22.8	23.22	24.18	24.36
25.22	24.92	23.4	23.68
24.4	23.96	23.14	23.14
24.14	24.08	23.574	23.96
23.38	23.48	21.9	22.68
23.18	22.7	26.62	26.24
23.7	23.72	21.96	22.66
24.32	24.02	24.42	24.58
22.96	22.6	23.02	22.66
22.9	22.98	24.56	24.08
24	24.34	23.1	23.92
23.42	23.44	23.7	24.14
23.8	23.58	23.34	23.6
24.5	24.4	24.96	25.12
23.82	23.64	23.14	22.8
22.62	22.68	23.22	23.64
22.7	22.54	22.62	22.36
22.34	21.86	21.8	21.6
23.04	23.14	20.44	20.2
22.44	22.56	23.04	23.2
21.32	21.54	21.52	21.1
21.58	21.42	21.68	21.82
22.1	22.28	21.9	21.7
22.74	22.66	21.78	21.96
22.6	22.54	20.62	20.66
21.08	21.04	21.64	21.42
20.86	20.6	21.52	21.36
		22.84	22.7
		22.4	22.48
		22.36	22.7
		21.88	22.06

Displaying Numerical Data

Frequency Table:

Interval	Freq	%
(20.5,21]	1	3.33
(21,21.5]	2	6.67
(21.5,22]	1	3.33
(22,22.5]	3	10.00
(22.5,23]	7	23.33
(23,23.5]	5	16.67
(23.5,24]	5	16.67
(24,24.5]	5	16.67
(24.5,25]	0	0.00
(25,25.5]	1	3.33

Histogram:

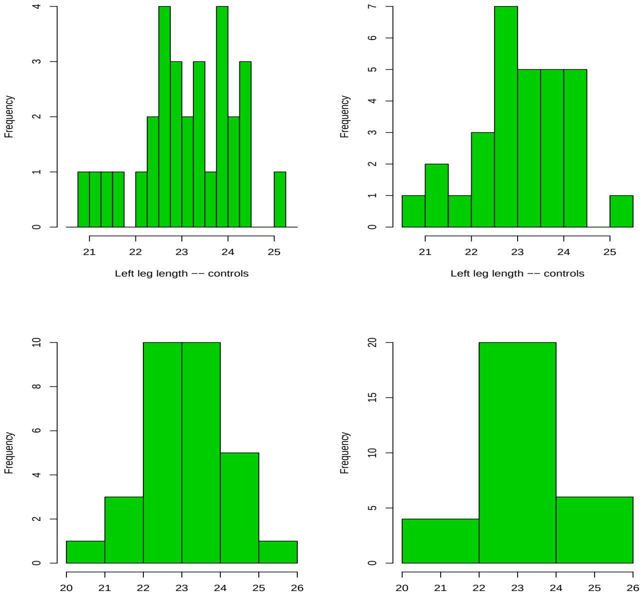


Displaying Numerical Data

Comments:

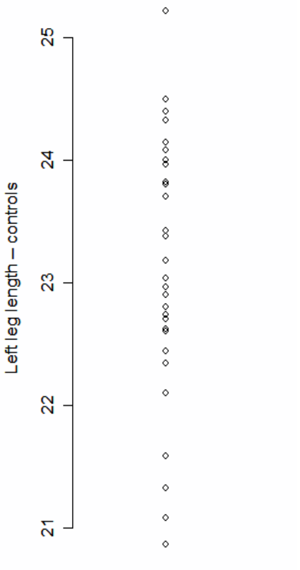
- If the frequency table or histogram is expressing fraction or percent of total, need to decide whether to include missing values in the denominator
- The AREA of each bar in a **histogram** is proportional to the frequency, so these are easiest to understand when the boundaries of the **bins** (that is, the intervals) are equal.
- If using software to create frequency tables or histograms, be sure to check what convention is used on the boundaries (e.g., is an interval defined $2 < x \leq 3$ or $2 \leq x < 3$).
- Appearance of a histogram can be quite difference depending on the number of groups used
- Note that for categorical data, we can use **bar plots**, which are very similar in appearance to histograms.

Displaying Numerical Data



Displaying Numerical Data

Dot Diagram:



Stem and Leaf Plot:

```
The decimal point
is at the |

  20 | 9
  21 | 136
  22 | 134667789
  23 | 002444788
  24 | 0011345
  25 | 2
```

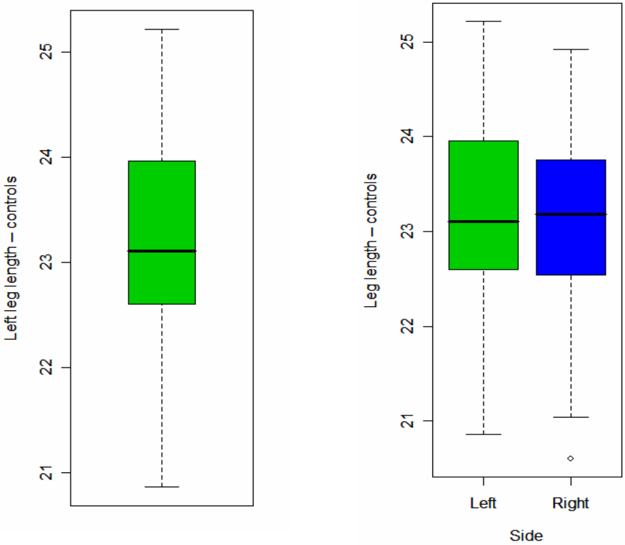
Displaying Numerical Data

Comments:

- **Dot plots** are very visual, but are not very useful if there are a large number of observations.
- **Stem and leaf plots** provide a relatively compact method of presenting all of the raw data; they are most useful for "moderate" amounts of data.
  - Constructed by dividing each measurement into a "stem" by truncating the observation in some position (e.g., tens digits), and the remaining value is the "leaf"
  - Leaves are ordered from smallest to biggest
  - Be careful with negative values!
  - Similar graphical appearance to a histogram, but displays more information

Displaying Numerical Data

Box-plot:



Displaying Numerical Data

Comments:

A **boxplot** is a simple display of the shape of a variable's distribution. Typically, a **five point** (lower quartile, median, upper quartile, plus "fences") summary is used, and often outlying observations are included. The exact form varies from package to package; in R, we have:

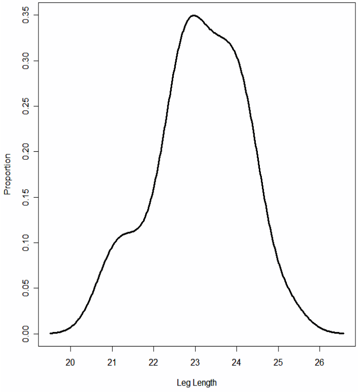
- The **median** (horizontal line)
- The **box** (the lower and upper quartiles, or **hinges**)
- The **whiskers**
- The **fences** (lower and upper horizontal lines, the smallest and largest values that are within 1.5 box-lengths of the box)
- **outliers** (plotted as circles, more than 1.5 box lengths above/below the box)

Some packages (e.g., SPSS) plot **extreme values** ($>$3 box lengths from box) as asterisks.

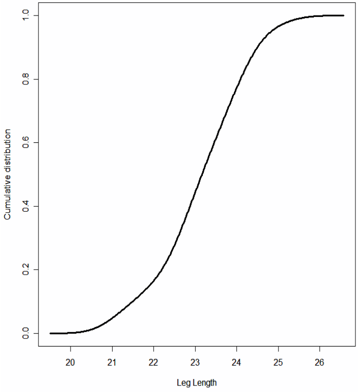
Displaying Numerical Data

Erica E. M. Moodie

Probability Density:



Cumulative Distribution:



Displaying Numerical Data

Comments:

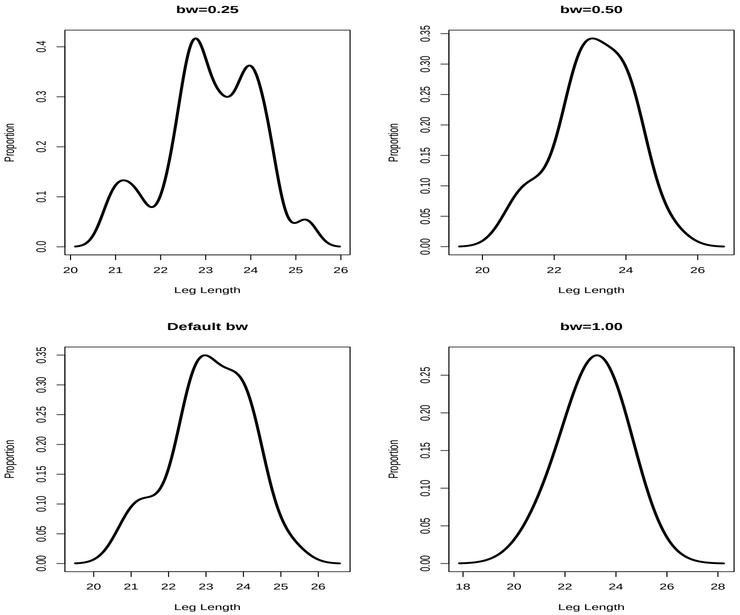
A **density curve** is like a smooth approximation of a histogram.

- The area under the curve is always equal to 1
- The y-axis is always non-negative
- We can use a density curve to (informally) assess whether a distribution appears to be **symmetric** or not; if a distribution is asymmetric, then it must have a **long left/right tail**.
- The smoothness of the density can be controlled by the bandwidth (how far away measurements can be from x and still influence density at x)

A **cumulative distribution** plots on the y-axis the proportion of values that are *less than or equal* to x.

Displaying Numerical Data

Erica E. M. Moodie



Summarizing Numerical Data

We wish to summarize the data in order to convey general trends or features that are present in the sample. Secondly, in order to propose an appropriate probability model, we want to match features in the sample data to features of one of the conventional probability distributions that may be used in more formal analyses. The principal features that we need to assess in the data sample are

1. The **location**, or "central tendency" in the sample.
2. The **mode**, or "most common" value in the sample.
3. The **scale** or **spread** in the sample.

These features of the sample are important because we can relate them **directly** to features of probability distributions.

Numerical Summaries: location

- The **sample mean**:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - The "point of balance" for the distribution
  - Defined only for numeric or binary data (not categorical); censored data may cause trouble in interpretation
  - The mean heavily influenced by large outliers, so this may not adequately reflect the a "typical value"
  - Can use the mean to compare distributions

Numerical Summaries: location

- The **sample median**:

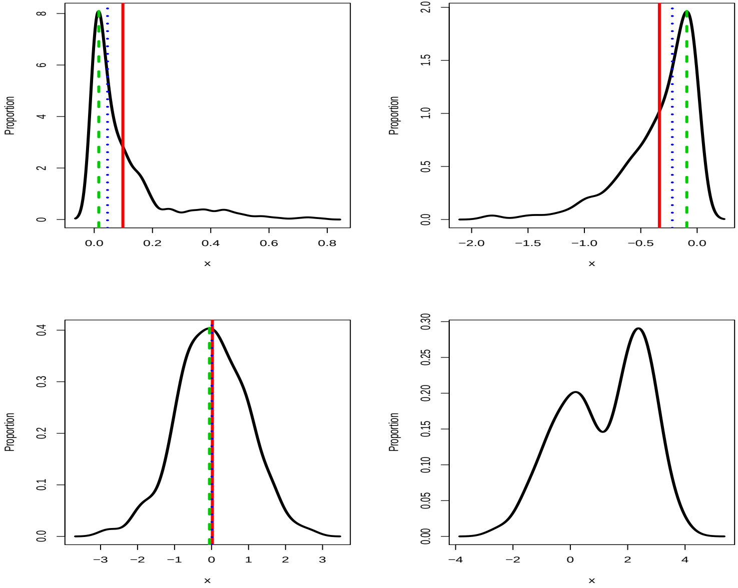
$$Q2 = \begin{cases} x_{[n/2]} \text{ if } n \text{ even} \\ \\ .5(x_{[(n-1)/2]} + x_{[(n+1)/2]}) \text{ if } n \text{ odd} \end{cases}$$

  - The value that is larger than half the sample and smaller than half the sample
  - Defined for any ordered variable (even categorical)
  - Sometimes can be used for censored data
  - Not sensitive to outlying values; more efficient for skewed data

Numerical Summaries: location

- The **sample mode**: the "most common" x.
  - Can use this for discrete data - definition: the most frequently observed value
  - ...or for continuous data - definition: the (local) maximum density; determined by a histogram or a probability density plot
  - For descriptive statistics: describes a common or typical values
  - For hypothesis generating: multi-modal distributions might indicate a mixture of populations

Numerical Summaries: location



Numerical Summaries: location

Erica E. M. Moodie

Sorted left leg lengths of controls birds:
20.86 21.08 21.32 21.58 22.10 22.34 22.44 22.60 22.62 22.70
22.74 22.80 22.90 22.96 **23.04 23.18** 23.38 23.38 23.42 23.70
23.80 23.82 23.96 24.00 24.08 24.14 24.32 24.40 24.50 25.22.

Sample mean: 23.11
Sample median: 23.11
Sample mode: 22.98

Numerical Summaries: spread

Erica E. M. Moodie

- The **range**: the lowest to the highest value
  - Only makes sense for ordered variables
- The **IQR** (inter-quartile range): the 25^{th} to 75^{th} percentile, or Q1 to Q3
  - Only makes sense for ordered variables
  - Contains the central 50% of the data
  - Less sensitive to outliers

Numerical Summaries: spread

Erica E. M. Moodie

- The **sample variance** and **standard deviation**:

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \qquad s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$$

  - Average squared distance from the mean
  - Useful for numerical variables (not categories)
  - Characterizes a distribution, can be used to compare distributions or to assess the validity of the assumptions for some statistical tests
  - Is more sensitive to outliers than the IQR
  - Has a role in the sampling distribution of the mean
  - Has a role in linear regression (squared error loss)

Numerical Summaries: spread

- The **sample variance** and **standard deviation**:

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \qquad s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$$

  - s^2 is an unbiased estimator of the population variance, σ^2; s is **not** an unbiased estimator of the population SD, σ
  - The SD is measured in the same units as the mean
  - Chebyshev's inequality: for any distribution that has a variance, **at least 90% of the data lie within 3 SDs of the mean**
  - For **Normal** data: (i) approx 67% of data like within 1 SD of the mean, and (ii) approx 95% of data like within 2 SDs of the mean

Numerical Summaries: spread

Why divide by $n-1$?

- The parameter that we are trying to estimate is the *population variance* (or standard deviation), σ^2, which is the average squared deviation from the true population mean, μ: $\sigma^2 = \text{Average}((x-\mu)^2)$.
- **IF** we knew μ, we could calculate $(x-\mu)^2$ for all the observed values x in our sample... but we don't know μ so we have to use our "best guess:" the sample mean, $\bar{x}$.
- If $\bar{x} < \mu$, then each x in our sample is likely to be smaller than μ and each squared deviation, $(x-\bar{x})^2$ tends to be smaller than $(x-\mu)^2$.
- Similarly, if $\bar{x} > \mu$, then each x in our sample is likely to be bigger than μ and each squared deviation, $(x-\bar{x})^2$ tends to be smaller than $(x-\mu)^2$.
- Therefore, $(x-\bar{x})^2$ tends to be smaller than $(x-\mu)^2$, and so we adjust the denominator of the variance formula to account for this.

Numerical Summaries: spread

Erica E. M. Moodie

Why divide by $n-1$? Let's look at this in more detail:

$$\begin{aligned}
\sum(x-\bar{x})^2 &= \sum(x-\mu+\mu-\bar{x})^2 \\
&= \sum(x-\mu)^2 - \sum(\bar{x}-\mu)^2 \\
&= \sum(x-\mu)^2 - n(\bar{x}-\mu)^2
\end{aligned}$$

so $\sum(x-\bar{x})^2 = \sum(x-\mu)^2 - n(\bar{x}-\mu)^2$.

Numerical Summaries: spread

Erica E. M. Moodie

If we used $\frac{1}{n}\sum(x-\bar{x})^2$, then the **average** we would observe (if we repeated this over a large number of sample – statistically speaking, the **expectation** of the quantity) is the average of $\frac{1}{n}\sum(x-\mu)^2 - (\bar{x}-\mu)^2$, which is $\sigma^2 - \frac{1}{n}\sigma^2 = \frac{n-1}{n}\sigma^2$.

- That is, on average, we under-estimate σ^2 by a factor or $1/n$ if we use n as the divisor when calculating the sample variance.
- Dividing instead by $n-1$ corrects this, so the average of all possible s^2 values is σ^2 (we have an **unbiased** estimator).

Numerical Summaries: spread

Suppose X takes on values 1,3,5 with probabilities $1/3$ each:

Erica E. M. Moodie

X:	1	3	5	$\mu = 3$
Prob(x):	1/3	1/3	1/3	
X-μ:	-2	0	2	
$(X-\mu)^2$:	4	0	4	σ^2: 8/3

Numerical Summaries: spread

Then there are 9 possible sample to consider, each equally likely:

		x_2		
		1	3	5
	1	$(1,1)$	$(1,3)$	$(1,5)$
x_1	3	$(3,1)$	$(3,3)$	$(3,5)$
	5	$(5,1)$	$(5,3)$	$(5,5)$

This gives rise to 9 equally likely sample means:

		x_2		
		1	3	5
	1	1	2	3
x_1	3	2	3	4
	5	3	4	5

Numerical Summaries: spread

Erica E. M. Moodie

If we use n as the divisor in calculating the variance we get:

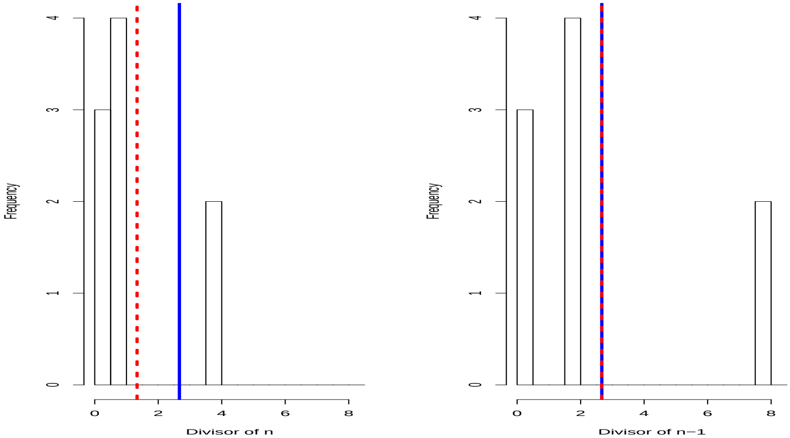
		x_2		
		1	3	5
	1	0	1	4
x_1	3	1	0	1
	5	4	1	0

If we use $n-1$ as the divisor in calculating the variance we get:

		x_2		
		1	3	5
	1	0	2	8
x_1	3	2	0	2
	5	8	2	0

Numerical Summaries: spread

Recall that the true variance is 8/3 (solid blue line). The dashed red line indicates the average of the sample variances:



Numerical Summaries: spread

Erica E. M. Moodie

Sorted left leg lengths of controls birds:
20.86 21.08 21.32 21.58 22.10 22.34 22.44 22.60 22.62 22.70
22.74 22.80 22.90 22.96 23.04 23.18 23.38 23.38 23.42 23.70
23.80 23.82 23.96 24.00 24.08 24.14 24.32 24.40 24.50 25.22.

Range: 20.86, 25.22
IQR: 22.60, 23.96 *or* 22.61, 23.93
Standard deviation:

$$\sqrt{\frac{(20.86-23.11)^2+(21.08-23.11)^2+...+(25.22-23.11)^2}{30-1}}$$

$$=\sqrt{32.36679/29}=1.056$$

Numerical Summaries: relative spread

- A measure of relative spread is the **coefficient of variation**:

$$CV = 100 * (s/\bar{x})\%$$

  - If CV is small, the spread of the data relative to the mean is small.
  - Note that this is not a very stable (and therefore not a useful) measure when the mean is near zero.

Reporting summaries

Erica E. M. Moodie

It is common to report a sample mean and variance, $\bar{x}$,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

and, in addition, a **standard error of the mean**

$$SE = \frac{s}{\sqrt{n}}.$$

But

- what is this quantity?
- why this formula?
- what if the data are **proportions**, or **counts out of** m**?**

Reporting summaries

Erica E. M. Moodie

For proportions, with x positive results out of n, then the estimate of the proportion is

$$\frac{x}{n}$$

and the **standard error** of this estimate is

$$\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}} = \sqrt{\frac{x\,(n-x)}{n^3}}$$

Reporting summaries

It is common to report

$$\bar{x} \pm SE$$

as a sample summary. However, it might be more appropriate to report a **confidence interval**

$$\bar{x} \pm 1.96 \times SE$$

- when is this formula valid?
- why this formula?

To answer these questions, some results from probability theory are needed.

Note that some authors write $\bar{x} \pm SD$ (that is, plus or minus the variable's standard deviation). This is bad practice! If you want to report the SD, be explicit and don't use "$\pm$": for example, write $\bar{x}(SD) = 2.3(0.89)$.

Relocating and rescaling numerical data

Erica E. M. Moodie

Suppose we have a variable X with mean μ_X and variance σ_X^2. If we transform X **linearly** (by adding a constant to X or by multiplying X by a constant) to get Y, what will be μ_Y, σ_Y, and σ_Y^2?

Transform.	μ_Y	σ_X	σ_X^2
$Y = X + a$	$\mu_X + a$	σ_X	σ_X^2
$Y = bX$	$b\mu_X$	$b\sigma_X$	$b^2\sigma_X^2$
$Y = bX + a$	$b\mu_X + a$	$b\sigma_X$	$b^2\sigma_X^2$
$Y = b(X + a)$			

Relocating and rescaling numerical data

Why would we want to relocate or rescale a variable?

- To change units – e.g., from °C to °F, from lbs to kilos, etc.
- To **standardize** a variable. Standardization involves two steps:
  1. Subtract a constant (usually the mean), so take $a = -\mu_X$
  2. Divide by a constant (usually the standard deviation), so take $b = 1/\sigma_X$
- Standardization gives $Y = b(X + a) = (X - \mu_X)/\sigma_X$.

- Question: What are μ_Y, σ_Y, and σ_Y^2?

Other transformations of numerical data

It may be necessary or advantageous to consider data **transformations;**

- $y_i = \log_{10} x_i$
- $y_i = \log x_i = \ln x_i$
- $y_i = \sqrt{x_i} = x_i^{1/2}$
- $y_i = x_i^{\alpha}$ some α
- $y_i = \log\left(\frac{x_i}{1 - x_i}\right)$

NOTE: Using a transformation is not any form of statistical trickery, but may be necessary to allow formal statistical assessment. For example, some statistical tests are only appropriate for symmetric distributions; if x is not symmetric, but a transformation of $y = f(x)$ is, we can instead use y for our formal tests.

How should you approach an analysis?

Erica E. M. Moodie

Get to know your data!

- What variables do you have?
- Do the values make sense?
- Can you answer the scientific question of interest with the data that you have?
- How will you handle missing data?
- Do any variables need to be transformed?
- How can you best summarize the sample ("Table 1")

How should you (start to) present an analysis?

Erica E. M. Moodie

Stroke-unit care for acute stroke patients: an observational follow-up study *Lancet* 2007; 369: 299–305

	Patients in stroke unit (n=4936)	Patients in control wards (n=6636)	Intra-class correlation coefficient
Age (years)	72 (12·9)	76 (12·2)	0·038
Men	2590 (52%)	3195 (47%)	0·001
Admission within 6 h	1926 (39%)	2526 (36%)	0·168
Intracranial haemorrhage	412 (7%)	859 (13%)	0·214
Atrial fibrillation	794 (16%)	1280 (19%)	0·034
Systolic blood pressure (mm Hg)	159 (28·9)	164 (37·4)	0·022
Diastolic blood pressure (mm Hg)	87 (14·4)	90 (14·4)	0·043
Unconsciousness	675 (13%)	1303 (20%)	0·034
Unconsciousness or motor impairment	3297 (70%)	4576 (69%)	0·066
Aphasia	1307 (25%)	1819 (26%)	0·042
Length of stay in hospital (days)	12 (11·3)	12 (12·2)	0·070

*Adjusted for hospital clusters. Data are mean (SD) or number (%).

Table 1: Distribution of baseline characteristics

How should you approach an analysis?

Erica E. M. Moodie

Are football referees really biased and inconsistent?: evidence on the incidence of disciplinary sanction in the English Premier League *J. R. Statist. Soc.* A (2007)

Table 1. Observed numbers of yellow cards incurred by the home and away teams, English Premier League, seasons 1996–1997 to 2002–2003†

Home team	*Distribution for the following numbers for away teams:*								*Total*
	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	
0	189	254	158	86	35	9	1	0	732
1	110	260	264	147	66	23	6	1	877
2	64	162	158	126	47	25	6	1	589
3	18	77	96	72	39	14	3	4	323
4	3	13	29	32	16	8	2	0	103
5	1	3	12	11	2	1	0	1	31
6	0	0	1	2	1	1	0	0	5
Total	385	769	718	476	206	81	18	7	2660

†Source: the Football Association.

How should you approach an analysis?

Table 1. Age-Adjusted Baseline Characteristics According to Intakes of Total Calcium and Total Vitamin D in the Women's Health Study

Characteristic	Calcium Intake Q1	Calcium Intake Q2	Calcium Intake Q3	Calcium Intake Q4	Calcium Intake Q5	*P* Value for Trend	Vitamin D Intake* Q1	Vitamin D Intake* Q2	Vitamin D Intake* Q3	Vitamin D Intake* Q4	Vitamin D Intake* Q5	*P* Val for Tren
No. of participants	6298	6298	6297	6297	6297		6298	6298	6298	6296	6297	
Mean age, y	54.5	54.7	54.9	55.4	56.4	<.001	54.3	54.7	55.3	55.4	56.2	<.00
Mean BMI	26.2	26.3	26.0	25.7	25.2	<.001	26.1	26.1	26.0	25.8	25.4	<.00
History of breast cancer in mother or sister, %	6.4	6.6	6.0	6.3	6.6	.79	6.6	6.6	6.7	6.6	5.6	.07
History of benign breast disease, %	30.3	32.1	32.0	32.6	36.1	<.001	31.6	31.7	33.4	32.1	34.3	.00
Mammogram screening, %†	51.3	57.5	60.5	62.9	67.2	<.001	54.6	58.6	60.4	61.6	64.4	<.00
Postmenopausal, %	66.7	65.5	65.8	65.8	68.5	.02	66.7	66.7	65.0	65.9	68.0	.59
Current users of postmenopausal hormone therapy, %	55.7	59.5	62.2	64.5	71.7	<.001	59.3	60.6	61.5	64.9	67.7	<.00
Current smokers, %	19.7	13.8	10.9	9.3	8.9	<.001	17.4	13.2	10.3	11.1	10.4	<.00
Current users of multivitamins, %	15.1	20.6	28.7	34.3	47.8	<.001	8.6	10.3	13.1	38.6	76.2	<.00
Calcium supplement users, %	5.7	16.3	36.1	56.2	89.5	<.001	23.6	26.0	29.8	48.8	75.9	<.00
Nulliparous women, %	12.6	13.2	13.3	13.4	14.9	.001	12.8	12.9	13.3	14.0	14.3	.00
Mean No. of children among parous women	3.0	2.9	2.9	2.9	2.8	<.001	2.9	2.9	2.9	2.9	2.8	.00
Mean age at first birth, y	24.5	24.6	24.7	24.9	24.8	<.001	24.6	24.6	24.8	24.8	24.6	.47
Mean age at menarche, y	12.5	12.4	12.4	12.4	12.4	.46	12.4	12.4	12.4	12.4	12.4	.20
Mean age at menopause, y	48.1	48.1	48.2	48.3	48.3	<.001	48.1	48.2	48.2	48.3	48.2	.05
Physical activity, kcal/wk	748	912	1016	1064	1126	<.001	807	906	978	1060	1123	<.00
Total calories intake, kcal/d	1630	1766	1785	1844	1623	<.001	1633	1754	1806	1865	1586	<.00
Alcohol intake, g/d	4.9	4.2	3.9	3.9	3.8	<.001	4.8	4.3	3.9	3.7	4.1	<.00
Total fat intake, g/d*	62	59	57	56	54	<.001	61	59	57	56	55	<.00
Phosphorus intake, mg/d*	1114	1254	1339	1431	1479	<.001	1131	1242	1350	1443	1451	<.00
Lactose intake, g/d*	6.3	11.6	16.0	21.2	23.3	<.001	7.3	12.2	17.1	21.9	19.9	<.00

How should you (start to) present an analysis?

- Always include some basic description of your sample, including how it was obtained
- (Almost) always include some basic tables and/or plots of your data
- Looking at the shape of a distribution (symmetric or not) will help you decide what information is most important (e.g., in HIV, viral load has a long right tail – mean is not very useful summary)
- ...but don't present everything you looked at, but do note anything unusual (e.g., long tail prompted you to use log transformation, etc.)

An introduction to the Normal (Gaussian) distribution

Erica E. M. Moodie

What is it?

- A distribution that describes continuous (numerical) data
- Can be used to approximate discrete data with enough categories to be considered nearly continuous
- Range is (technically) infinite, though the probability of seeing very large or very small values is extremely tiny
- Fully described by only two parameters, the mean and variance (μ and σ^2)
- Short-hand: $X \sim \mathcal{N}(\mu, \sigma^2)$

An introduction to the Normal (Gaussian) distribution

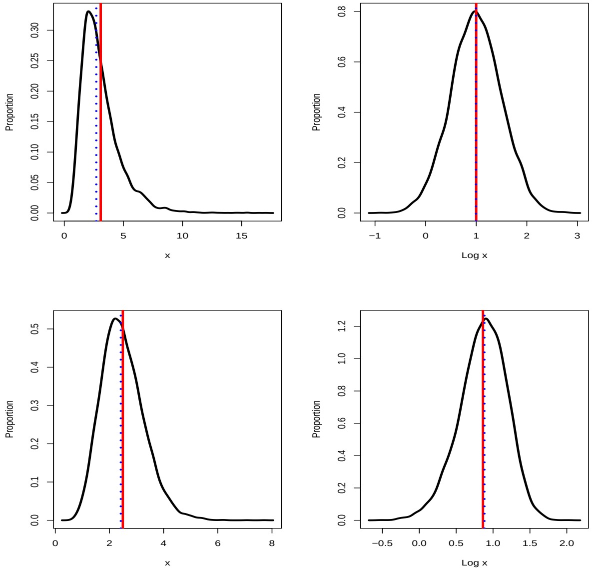
Where does Normal data come from?

- Natural processes
  - Blood pressure
  - Height
  - Weight
- “Man-made” (or derived)
  - Binomial (proportion) and Poisson (count) data are approximately Normal under certain conditions
  - Sums and means of random variables (Central Limit Theorem)
  - Data can sometimes be made to look Normal via transformations (squares, logs, etc)

An introduction to the Normal distribution

Erica E. M. Moodie

Transformations to symmetrize:



The Normal distribution

For Normal data, we can use the Gaussian tables to answer the questions:

- What is the probability that a single observation X is
  - greater than X^*?
  - less than X^*?
  - between X_L^* and X_U^*?
- That is, we can find out information about the percent distribution of X as a function of thresholds X^*.

- We can also use the Normal tables to find out information about thresholds X^* that will contain particular percentages of the data. I.e., we can find what threshold values will
  - Exclude the lower p*% of a population
  - Exclude the upper p*% of a population
  - Contain the middle p*% of a population

The Normal distribution

Erica E. M. Moodie

We can use the Gaussian tables to answer these questions **no matter what the values of** μ and σ^2.

That is, the % of the Normal distribution falling between $X_L^* = \mu + m_1\sigma$ and $X_U^* = \mu + m_2\sigma$ where m_1, m_2 are any multiples **remains the same** for any μ and σ^2.

How so??

Using Z, a **standardized** version of $X \sim \mathcal{N}(\mu, \sigma^2)$!

The Normal distribution

An illustration using IQ scores, which we presume have a $\mathcal{N}(100, 13)$ distribution of scores.

Q1: What percentage of scores are **above** 130?
Two steps:

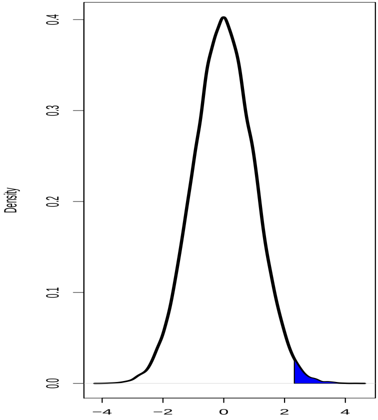
1. Change of location from $\mu_X = 100$ to $\mu_Z = 0$
2. Change of scale from $\sigma_X = 13$ to $\sigma_Z = 1$

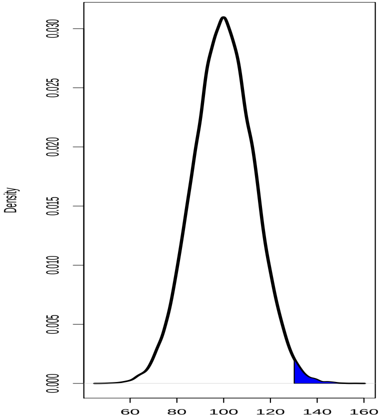
Together, this gives us

$$Z = \frac{X - \mu_X}{\sigma_X} = \frac{130 - 100}{13} = 2.31$$

The Normal distribution

The position of $X{=}130$ in a $\mathcal{N}(100, 13)$ distribution is the same as the place of $Z = 2.31$ on the $\mathcal{N}(0, 1)$, which we call the **standardized** Normal distribution (or Z-distribution).





The Normal distribution

Erica E. M. Moodie

(The percent above $X = 130$) = (% above $Z = 2.31$) =1.04%

How do we know this? We look at the lower tail probability of 2.31 [i.e., the % below 2.31], and then subtract it from 1:

1. $P(X < 130) = P(Z < 2.31) = 0.9896$
2. $P(X > 130) = 1 - P(X < 130) = 0.0104$

So 130 is the 98.96^{th} percentile.

The Normal distribution

Q2: What is the 75^{th} percentile of the IQ scores distribution?
We now have to reverse the sequence of steps:

- Start of probability 0.75 in the body of the table; this corresponds to a z value of 0.675.
- The z value is from a N(0,1) distribution, so we need to convert this to the IQ scale of a N(100,13):
  1. 0.675 SDs on z scale $=$ 0675$\times$13 SDs on X scale $=$ 8.8 IQ points
  2. $X = \mu_X + z\sigma_X = 100 + 8.8 = 108.8$

This gives us that 75% of the IQ scores fall below 108.8.

Table entry for z is the area under the standard normal curve to the left of z.

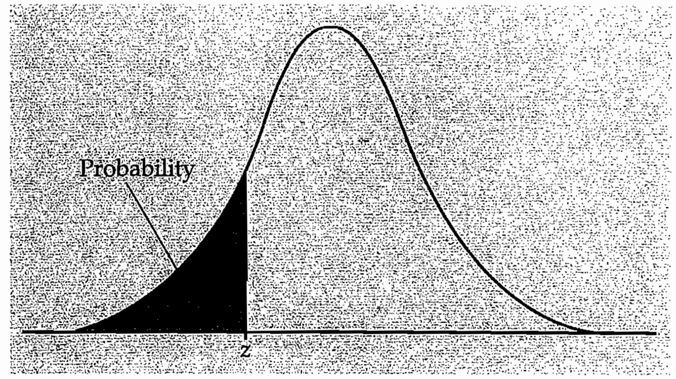


TABLE A Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

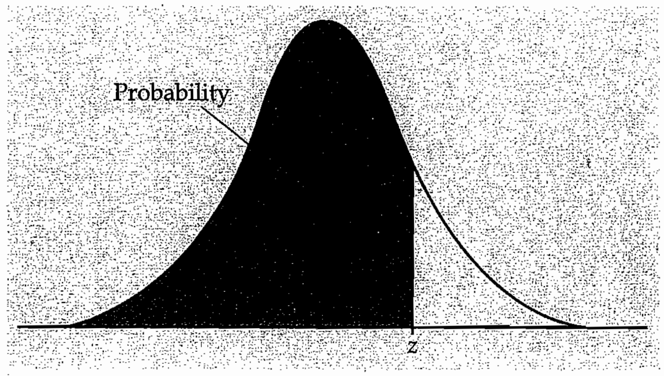


Table entry for z is the area under the standard normal curve to the left of z.

TABLE A Standard normal probabilities (continued)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998