Introduction to Statistics

Source

Material, examples and data sets come from:

Open Intro Statistics, Second edition
by
D.M.Diez, C.D.Barr and M. Cetinkaya-Rundel

https://www.openintro.org/

Motivation

Case Study:

Q: Does the use of stents reduce the risk of stroke?

Motivation

Treatment group. Patients in the treatment group received a stent and medical management. The medical management included medications, management of risk factors, and help in lifestyle modification.

Control group. Patients in the control group received the same medical management as the treatment group, but they did not receive stents.

Motivation

Results from the stent study

Patient	group	0-30 days	0-365 days
1	treatment	no event	no event
2	treatment	stroke	stroke
3	treatment	no event	no event
⋮	⋮	⋮	
450	control	no event	no event
451	control	no event	no event

Motivation

Summary of the stent study

	0-30 days		0-365 days	
	stroke	no event	stroke	no event
treatment	33	191	45	179
control	13	214	28	199
Total	46	405	73	378

- Proportion who had a stroke in the treatment (stent) group: $45/224 = 0.20 = 20\%$.
- Proportion who had a stroke in the control group: $28/227 = 0.12 = 12\%$.

Motivation

- An additional 8% of patients in the treatment group had a stroke
- This is contrary to what doctors expected, i.e. the stents would reduce the rate of strokes.
- Statistical question: do the data show a ”real” difference between the groups?

Basics

Statistics is the study of how best to collect, analyse and draw conclusions from data.

Basics

- **Process of investigation**
  - Identify a question or problem
  - Collect relevant data on the topic
  - Analyse the data
  - Form a conclusion

Examples

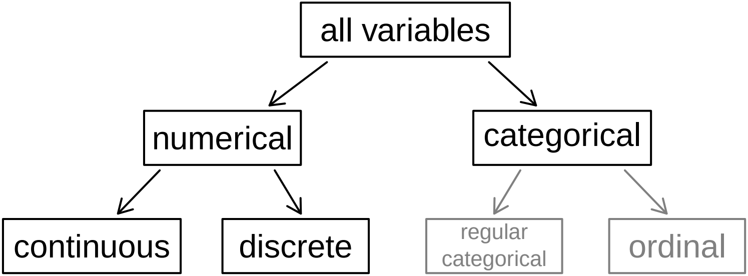
Data Matrix

	spam	num_ char	line_ breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	no	15,829	242	html	small

Variable description

variable	**description**
spam	Specifies whether the message was spam
num_char	The number of characters in the email
line_breaks	The number of line breaks in the email (not including text wrapping)
format	Indicates if the email contained special formatting, such as bolding, tables, or links, which would indicate the message is in HTML format
number	Indicates whether the email contained no number, a small number (under 1 million), or a large number

Types of Variables



Numerical variables: num_ char, line_ breaks

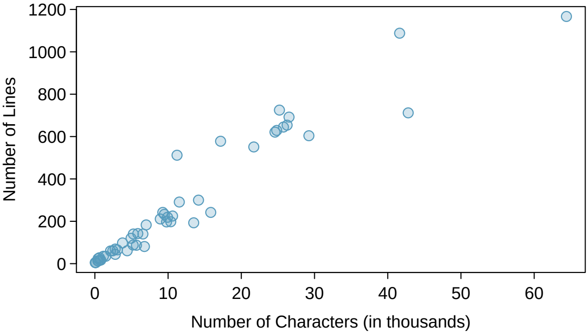
Categorical variables: spam, format, number

Exploring numerical data

1. Scatterplots for paired data
2. Dot plots and mean
3. Histograms and Shape
4. Variance and Standard deviation
5. Box plots, quartiles and median
6. Outliers and robust statistics

Scatterplots

A **scatterplot** provides a case-by-case view of data for two numerical variables.



Scatterplots are helpful in quickly spotting associations relating variables, whether those associations come in the form of simple trends or whether those relationships are more complex

Scatterplots-Association

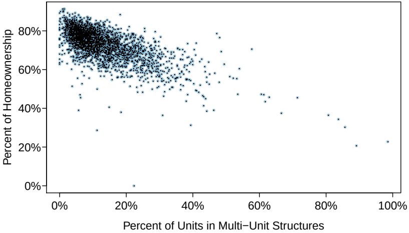


Figure: Negative association

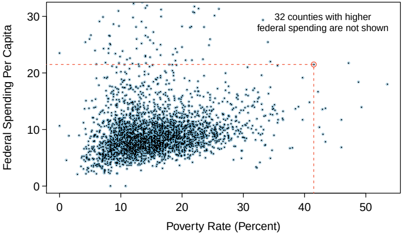


Figure: Positive association

Scatterplots- Trends

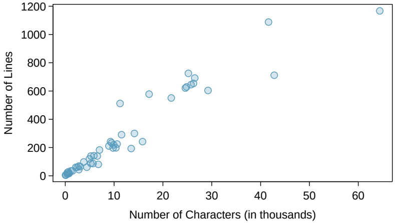


Figure: Linear trend

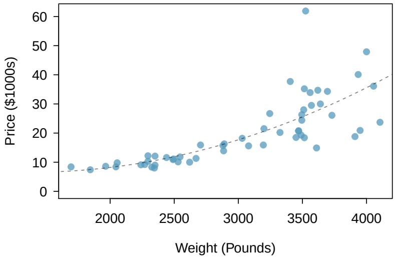
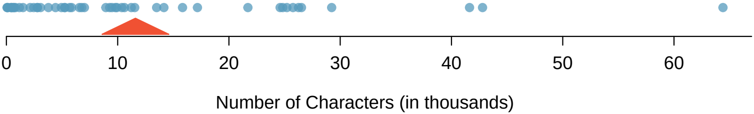


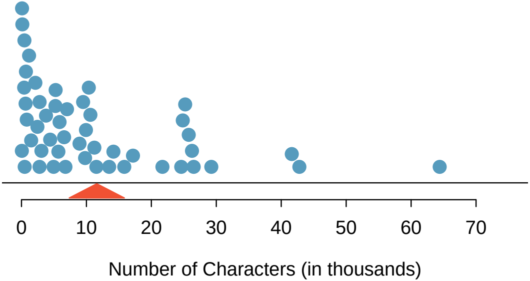
Figure: Non-Linear trend

Dot plots

A **dot plot** is a one-variable scatterplot.



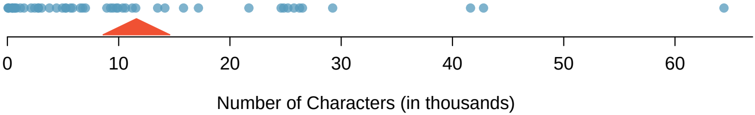
Stacked dot plot



Mean

The **mean**, sometimes called the **average**, is a common way to measure the centre of a **distribution** of data. To find the mean number of characters in the 50 emails, we add up all the character counts and divide by the number of emails.

$$\bar{x} = \frac{21.7 + 7.0 + \cdots + 15.8}{50} = 11.6$$



Mean

The **sample mean** of a numerical variable is computed as the sum of all of the observations divided by the number of observations:

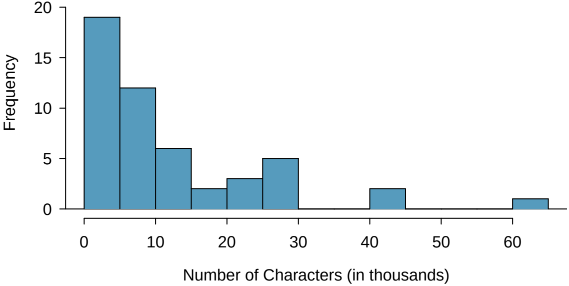
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

where $x_1, x_2, \ldots, x_n$ represent the n observed values.

Histograms

Bins/Frequency

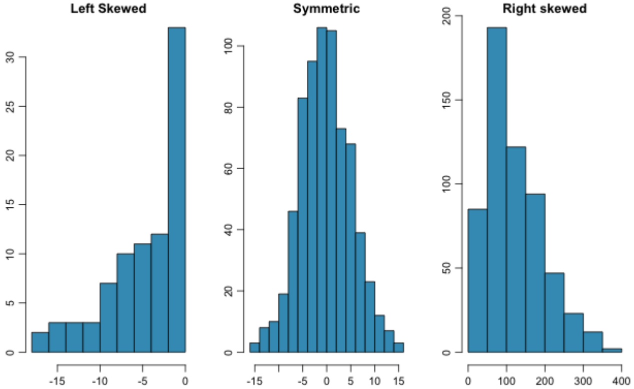
Characters (in thousands)	0-5	5-10	10-15	15-20	20-25	· · ·	60-65
Count	19	12	6	2	3	· · ·	1



Characteristics of Histograms-Skew

Skew: When data trail off in one direction, the distribution has a **long tail**.

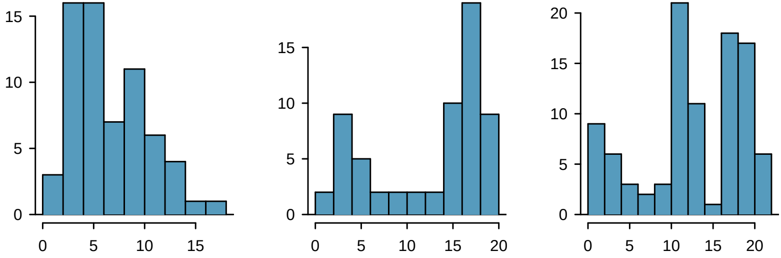
- If a distribution has a long left tail, it is **left skewed**.
- If a distribution has a long right tail, it is **right skewed**.
- Data sets that show roughly equal trailing off in both directions are called **symmetric**.



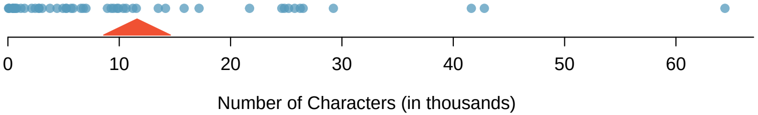
Characteristics of Histograms-Modes

Modes: A **mode** is represented by a prominent peak in the distribution.

- **unimodal**
- **bimodal**
- **multimodal**



Deviation



Variance and standard deviation

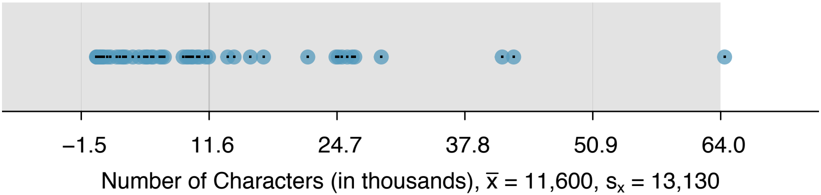
Deviation of observation i is the distance of an observation from its mean, $x_i - \bar{x}$.

The **sample variance**, s^2, of a numerical variable is computed as the sum of the squared deviation all of the observations divided by the number of observations minus one:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

where $x_1, x_2, \ldots, x_n$ represent the n observed values.

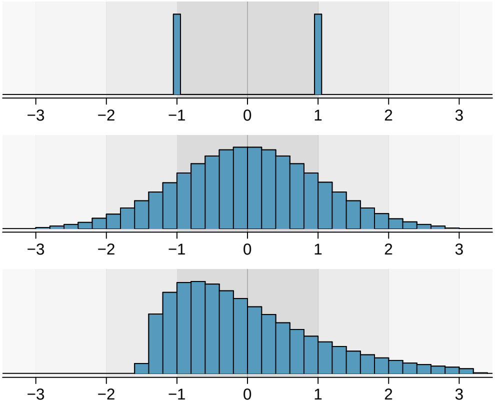
The **standard deviation**, s, is defined as the square root of the variance.



In the num_char data, 41 of the 50 emails (82%) are within 1 standard deviation of the mean, and 47 of the 50 emails (94%) are within 2 standard deviations.

Is it enough?

Three very different population distributions with the same sample mean $\bar{x} = 0$ and standard deviation $s = 1$.



Boxplot

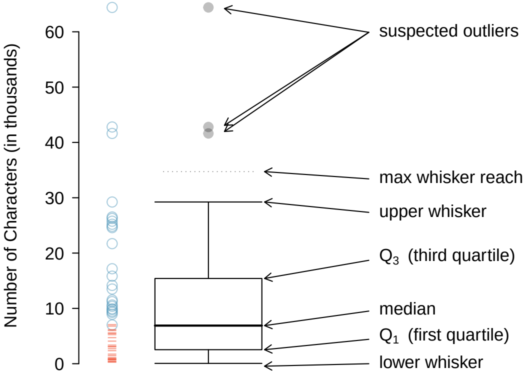


Figure: A vertical dot plot next to a labelled box plot for the number of characters in 50 emails. The median (6,890), splits the data into the bottom 50% and the top 50%, marked in the dot plot by horizontal dashes and open circles, respectively.

Median-IQR-Whiskers

If the data are ordered from smallest to largest, the **median** is the observation right in the middle. If there are an even number of observations, there will be two values in the middle, and the median is taken as their average.

The **interquartile range** (IQR) is the length of the box in a box plot. It is computed as

$$IQR = Q_3 - Q_1$$

where Q_1 and Q_3 are the 25^{th} and 75^{th} percentiles.

Whiskers attempt to capture the data outside of the box, however, their reach is never allowed to be more than $1.5 \times IQR$.

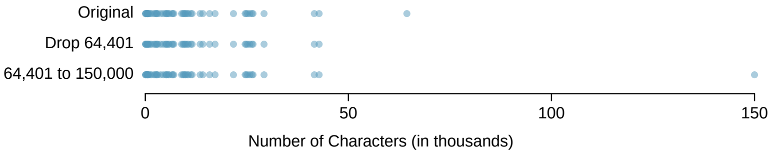
Outliers

An **outlier** is an observation that appears extreme relative to the rest of the data.

Examination of data for possible outliers serves many useful purposes, including

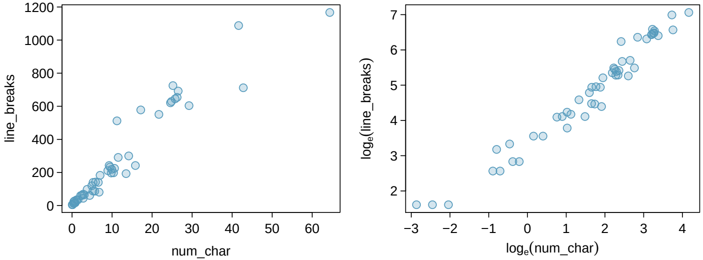
1. Identifying strong skew in the distribution.
2. Identifying data collection or entry errors. For instance, we re-examined the email purported to have 64,401 characters to ensure this value was accurate.
3. Providing insight into interesting properties of the data.

Robust statistics



	robust		**not robust**	
scenario	median	IQR	$\bar{x}$	s
original num_char data	6,890	12,875	11,600	13,130
drop 66,924 observation	6,768	11,702	10,521	10,798
move 66,924 to 150,000	6,890	12,875	13,310	22,434

Data Transformation



Common transformation functions

- $log(x)$
- $\sqrt{x}$
- $\frac{1}{x}$

Categorical Data I

Summary table

number			
none	small	big	Total
549	2827	545	3921

A table that summarises data for a categorical variable is called a **frequency table.**

number			
none	small	big	Total
$\frac{549}{3921} = 0.14$	$\frac{2827}{3921} = 0.72$	$\frac{545}{3921} = 0.14$	1.00

Barplot

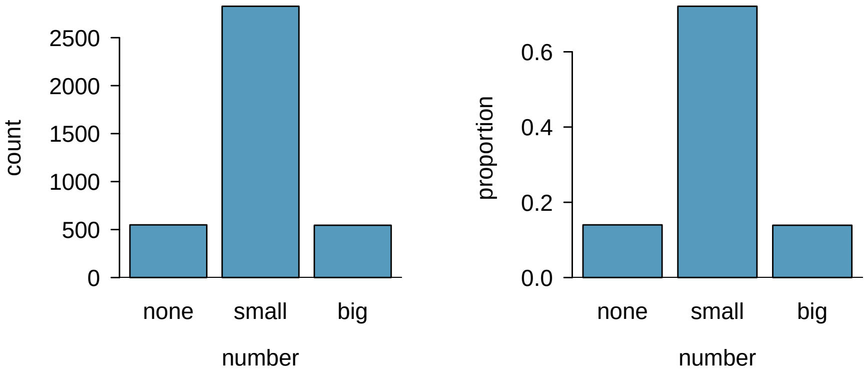


Figure: Two bar plots of number. The left panel shows the counts, and the right panel shows the proportions in each group.

Categorical Data II

A table that summarises data for two categorical variables in this way is called a **contingency table**.

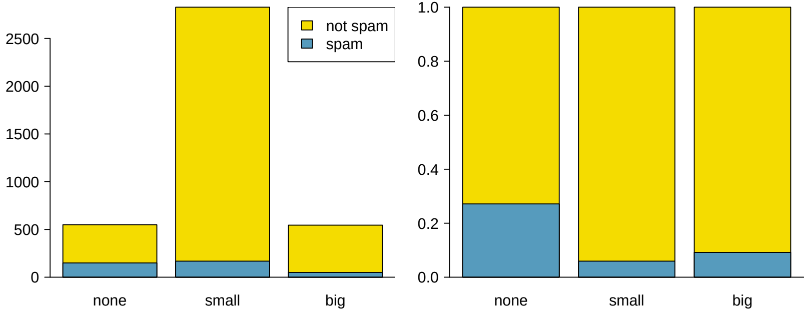
		number			
		none	small	big	Total
spam	spam	149	168	50	367
	not spam	400	2659	495	3554
	Total	549	2827	545	3921

A contingency table with row proportions

	none	small	big	Total
spam	$\frac{149}{367} = 0.406$	$\frac{168}{367} = 0.458$	$\frac{50}{367} = 0.136$	1.000
not spam	$\frac{400}{3554} = 0.113$	$\frac{2657}{3554} = 0.748$	$\frac{495}{3554} = 0.139$	1.000
Total	$\frac{549}{3921} = 0.140$	$\frac{2827}{3921} = 0.721$	$\frac{545}{392} = 0.139$	1.000

Segmented bar plot

A **segmented barplot** is a graphical display of contingency table information.



Categorical II

A contingency table with column proportions

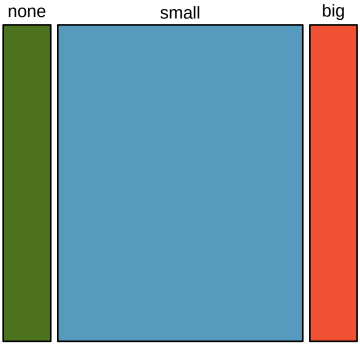
	none	small	big	Total
spam	$149/549 = 0.271$	$168/2827 = 0.059$	$50/545 = 0.092$	$367/3921 = 0.094$
not spam	$400/549 = 0.729$	$2659/2827 = 0.941$	$495/545 = 0.908$	$3684/3921 = 0.906$
Total	1.000	1.000	1.000	1.000

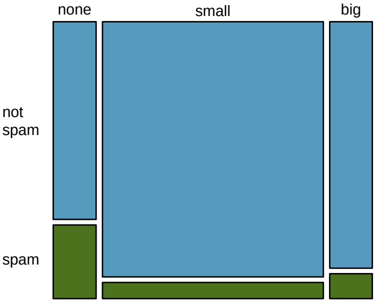
Mosaic plot

A mosaic plot is a graphical display that allows you to examine the relationship among two or more categorical variables.

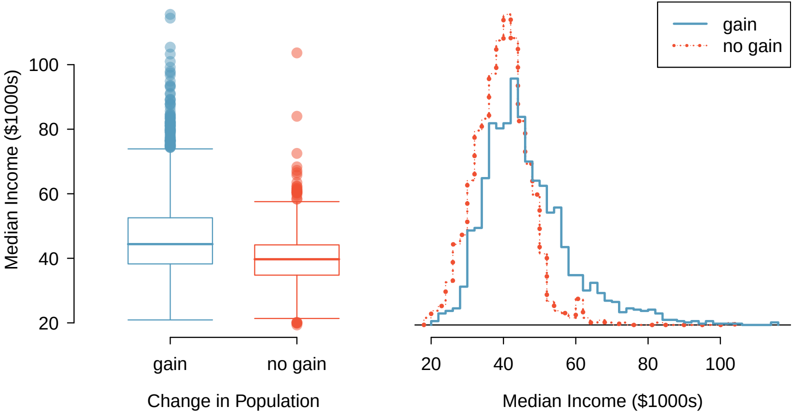
1. The mosaic plot is a square with length one.
2. The horizontal bars widths are proportional to the probabilities associated with the first categorical variable.
3. The vertical bars widths are proportional to the conditional probabilities of the second categorical variable.

Mosaic Plots





Comparing numerical data across groups

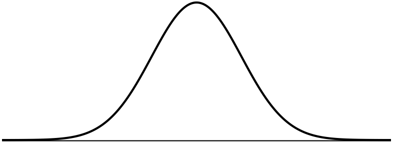


Some probability

Common Probability distributions

- Normal distribution
- Bernoulli
- Geometric distribution
- Binomial distribution

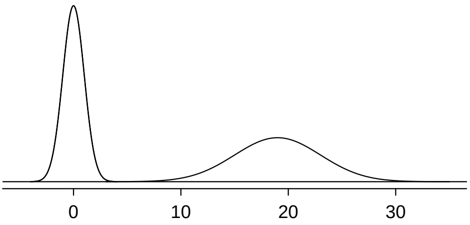
Normal Distribution



- $X \sim \mathcal{N}(\mu, \sigma^2)$
- Probability density function:
  $f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- Parameters:
  - μ denotes the mean
  - σ^2 denotes the variance[a]

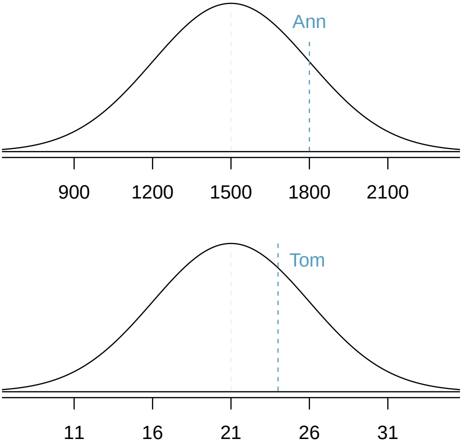
[a] σ denotes the standard deviation

Normal Distribution



$$N(\mu = 0, \sigma = 1) \quad \text{and} \quad N(\mu = 19, \sigma = 4)$$

Score comparison



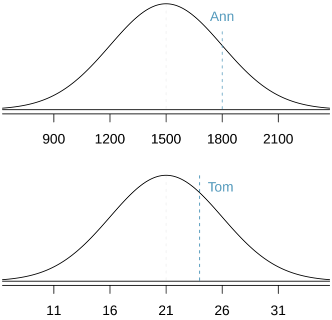
- SAT follow $\mathcal{N}(1500, 300)$
- ACT follow $\mathcal{N}(21, 5)$

Standardizing with Z scores

The **Z score** of an observation is the number of standard deviations it falls above or below the mean. We compute the Z score for an observation x that follows a distribution with mean μ and standard deviation σ using

$$Z = \frac{x - \mu}{\sigma}$$

Score comparison

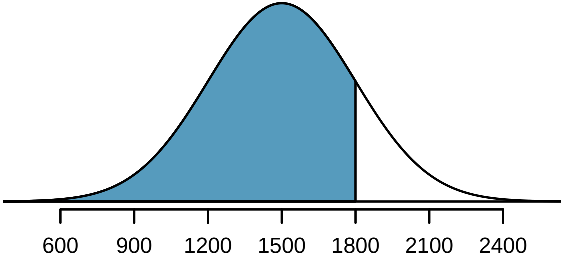


- $Z_{Ann} = \frac{x_{Ann} - \mu_{SAT}}{\sigma_{SAT}} = 1$

- $Z_{Tom} = \frac{x_{Tom} - \mu_{ACT}}{\sigma_{ACT}} = 0.6$

Percentile

- What percentile does Ann fall among all SAT test-takers?
- Ann's **percentile** is the percentage of people who earned a lower SAT score than her. Ann is in the $84th$ percentile of SAT takers.

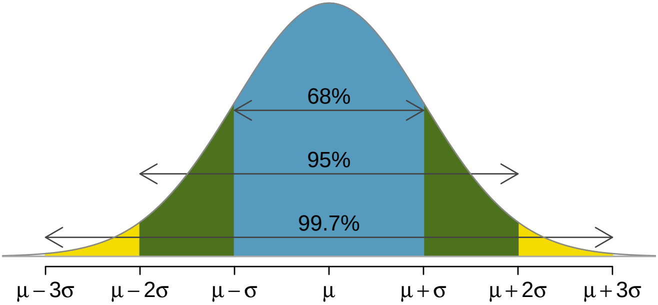


Normal probability table

	Second decimal place of Z									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	**0.8413**	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

68-95-99.7 Rule[1]

68% – 95% – 99.7% are, respectively, the probabilities for falling within 1, 2, and 3 standard deviations of the mean in a normal distribution.



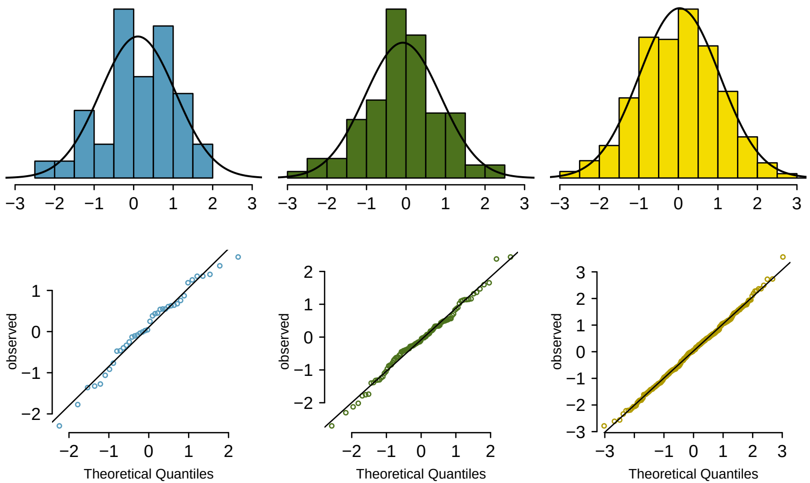
[1] Otherwise know as the **three-sigma rule of thumb**

Evaluating the normal approximation

- Best fitting normal curve
- Normal probability plot/ Quantile-quantile plot

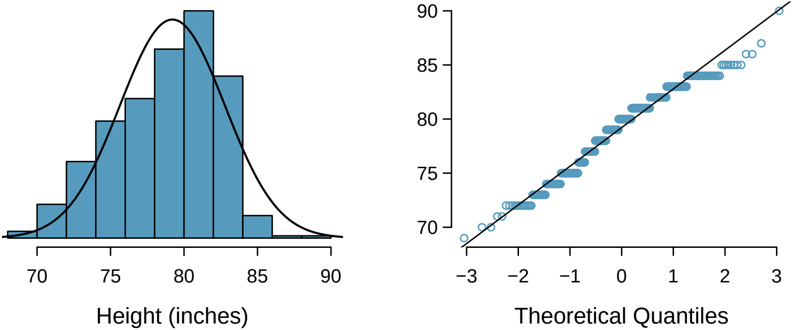
Simulated normal distribution

Histograms and normal probability plots for three simulated normal data sets; $n = 40$ (left), $n = 100$ (middle), $n = 400$ (right).



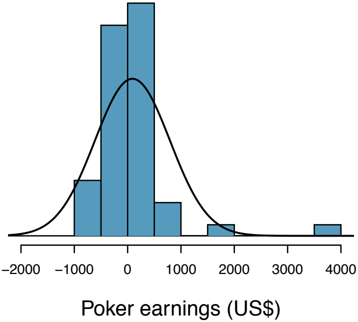
NBA's players height

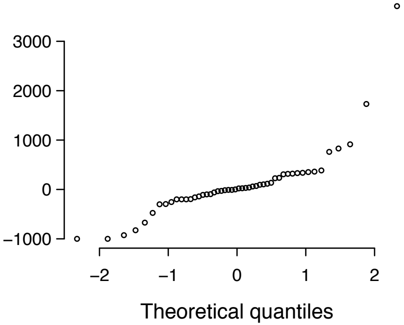
We consider all 435 NBA players from the 2008-9 season.



Poker earnings

We consider the poker earnings of an individual over 50 days. Can we approximate poker winnings by a normal distribution?





Bernoulli

Consider a coin toss with outcome $\{H, T\}$ and the random variable X :

$$X = \begin{cases} 1, & \text{for H (sucess)} \\ 0, & \text{for T (failure)} \end{cases}$$

Then

$$P(X = 1) = P(\text{sucess}) = p$$

and

$$P(X = 0) = P(\text{fail}) = 1 - p$$

Bernoulli

Bernoulli random variable

If X is a random variable that takes value 1 with probability of success p and 0 with probability $1-p$, then X is a Bernoulli random variable with mean and standard deviation

$$\mu = p \qquad\qquad \sigma = \sqrt{p(1-p)}$$

Geometric distribution

What is the probability of finding the first "success" after 4 trials?

$$\{T, T, T, H\} \Rightarrow (1-p)^3 p$$

Geometric distribution

Let X is a random variable that takes the value k, where k is number of trial needed to get the first "sucess", then

$$P(X = k) = (1-p)^{k-1} p$$

Additionally, the mean, variance, and standard deviation of the number of observed successes are

$$\mu = \frac{1}{p} \qquad \sigma^2 = \frac{1-p}{p^2} \qquad \sigma = \sqrt{\frac{1-p}{p^2}}$$

Binomial distribution

Consider 3 **coin tosses, what is the probability of** 2 **heads?**
Possible outcomes

$$(H, H, T) \quad (H, T, H) \quad (T, H, H),$$

with

$$P((H, H, T)) = p^2(1-p)$$
$$P((H, T, H)) = P((T, H, H)) = p^2(1-p).$$

Then

$$P(\text{'2 heads'}) = 3p^2(1-p) = \binom{3}{2} p^2(1-p)$$

Binomial distribution

Binomial distribution

Let X is a random variable that takes the value k, where k is the number of successes in n independent trials, then

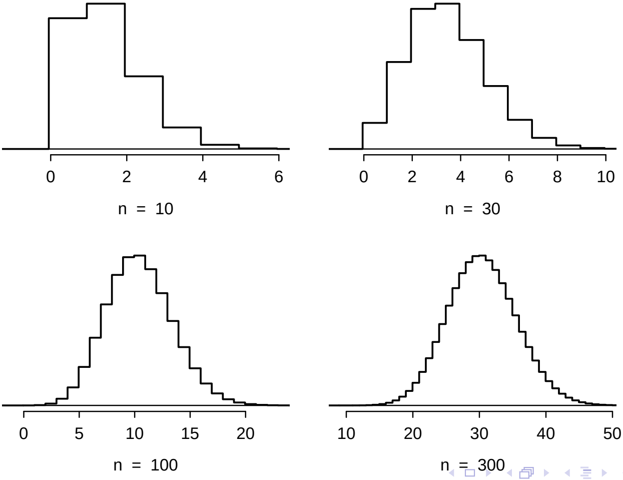
$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}=\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

Additionally, the mean, variance, and standard deviation of the number of observed successes are

$$\mu=np \qquad \sigma^2=np(1-p) \qquad \sigma=\sqrt{np(1-p)}$$

Normal approximation to the binomial distribution-Graph

Hollow histograms of samples from the binomial model when $p = 0.10$.



Normal approximation of the binomial distribution

Normal approximation of the binomial distribution
The binomial distribution with probability of success p is nearly normal when the sample size n is sufficiently large that

1. $np \geq 10$, and
2. $n(1-p) \geq 10$

The approximate normal distribution has parameters corresponding to the mean and standard deviation of the binomial distribution:

$$\mu = np \qquad\qquad \sigma = \sqrt{np(1-p)}$$

Foundations for inference

- A population
- Random sample
- **Statistical inference** is drawing conclusions based on the sample:
  1. point estimates; the value of a particular variable
  2. an interval estimate; establishing confidence intervals for point estimates
  3. Hypothesis testing

Inference on the population mean

Q: How sure are we that the estimated mean, $\bar{x}$, is near the true population mean, μ?

Data

Data in **run10**, all 16,924 runners who finished the 2012 Cherry Blossom 10 mile run in Washington, DC.

ID	time	age	gender	state
1	92.25	38.00	M	MD
2	106.35	33.00	M	DC
3	89.33	55.00	F	VA
4	113.50	24.00	F	VA
⋮	⋮	⋮	⋮	⋮
16923	122.87	37.00	F	VA
16924	93.30	27.00	F	DC

variable	**description**
time	Ten mile run time, in minutes
age	Age, in years
gender	Gender (M for male, F for female)
state	Home state (or country if not from the US)

Sample

run10Samp a simple random sample of 100 runners from the 2012 Cherry Blossom Run.

ID	time	age	gender	state
1983	88.31	59	M	MD
8192	100.67	32	M	VA
11020	109.52	33	F	VA
⋮	⋮	⋮	⋮	⋮
1287	89.49	26	M	DC

To estimate the average 10 mile run time of all participants, take the average time for the sample:

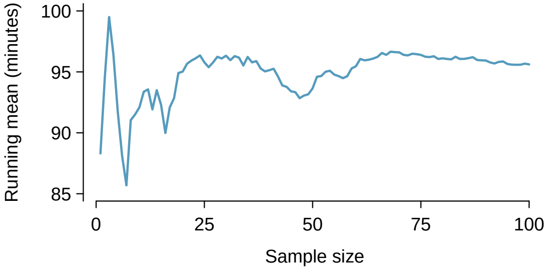
$$\bar{x} = \frac{88.22 + 100.58 + \cdots + 89.40}{100} = 95.61$$

Population parameters vs. Point estimates

time	estimate	parameter
mean	95.61	94.52
median	95.46	94.03
st. dev.	15.78	15.93

Running mean

A **running mean is** a sequence of means, where each mean uses one more observation in its calculation than the mean directly before it in the sequence.

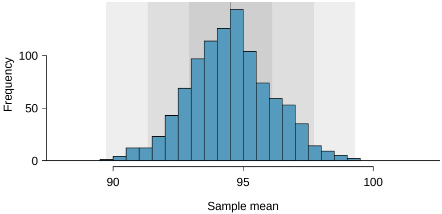


Note: The mean tends to approach the true population average as more data become available.

Sampling distribution

Take 1000 random samples of 100 individuals from **run10**, with estimated average time:

Sample	estimated average
1	95.61
2	93.43
3	93.43
⋮	
1000	94.16



Sampling distribution

- The sampling distribution represents the distribution of the point estimates based on samples of a fixed size from a certain population.
- Uncertainty about the estimate is captured by the standard deviation of the estimate, called the **standard error (SE)**.
- *Computing SE for the sample mean:* Given n independent observations from a population with standard deviation σ, the standard error of the sample mean is equal to

$$SE = \frac{\sigma}{\sqrt{n}} \tag{1}$$

Confidence interval

A plausible range of values for the population parameter is called a **confidence interval**.

An approximate 95% confidence interval

- The standard error represents the standard deviation associated with the estimate.
- Roughly 95% of the time the estimate will be within 2 standard errors of the parameter.
- If the interval spreads out 2 standard errors from the point estimate, we can be roughly 95% confident that we have captured the true parameter:

$$\text{point estimate} \ \pm \ 2 \times SE$$

A sampling distribution for the mean

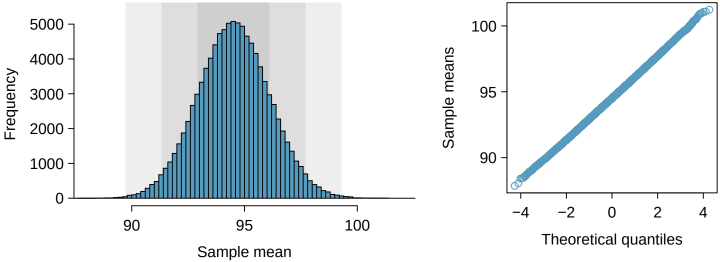
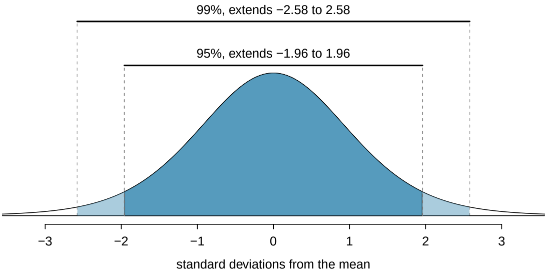


Figure: The left panel shows a histogram of the sample means for 100,000 different random samples. The right panel shows a normal probability plot of those sample means.

⇓

The distribution of the sample mean is well approximated by a normal model.

Confidence interval



- The 95% condidence interval for mean estimate is

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

- The 99% condidence interval for mean estimate is

$$\bar{x} \pm 2.58 \times SE_{\bar{x}}$$

Checking normality

Conditions for $\bar{x}$ being nearly normal and SE being accurate

Important conditions to help ensure the sampling distribution of $\bar{x}$ is nearly normal and the estimate of SE sufficiently accurate:

- The sample observations are independent.
- The sample size is large: $n \geq 30$ is a good rule of thumb.
- The population distribution is not strongly skewed. (We check this using the sample distribution as an estimate of the population distribution.)

Additionally, the larger the sample size, the more lenient we can be with the sample's skew.

Checking normality

Alternative conditions for normality
If the population of cases is known to be nearly normal and the population standard deviation σ is known, then the sample mean $\bar{x}$ will follow a nearly normal distribution $N(\mu, \sigma/\sqrt{n})$ if the sampled observations are also independent.

Confidence interval

Confidence interval for any confidence level

If the point estimate follows the normal model with standard error SE, then a confidence interval for the population parameter is

$$\text{point estimate } \pm \; z^{\star} SE$$

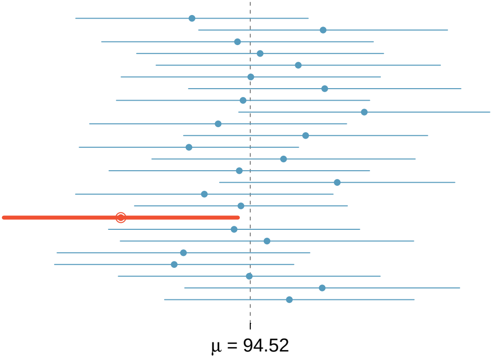
where $z^{\star}$ corresponds to the confidence level selected and the standard error is

$$SE = \begin{cases} \frac{\sigma}{\sqrt{n}}, & \text{if } \sigma \text{ is known} \\ \frac{s}{\sqrt{n}}, & \text{otherwise.} \end{cases}$$

Confidence interval

Q: What does 95% confident mean? If we take a number of samples and construct their confidence intervals, about 95% will include the actual mean μ.

Figure: Twenty-five samples of size $n = 100$ from the run10 data set and their respective confidence intervals.



Interpretation

Correct interpretation:
We are XX% confident that the population parameter is between...
Example: The 95% confidence interval for the time mean, base on run10Samp, is

$$(92.45, 98.77)$$

Hence, we are 95% confident that the population mean in between 92.45 and 98.77.

Hypothesis testing

Cherry Blossom Run

- The average time for all runners who finished the Cherry Blossom Run in 2006 was 93.29.
- The sample average time of the sample runners who finished the Cherry Blossom Run in 2012 was 95.61.
- Does the sample data from 2012 provides strong evidence that the runners were faster or slower than those in runners in 2006?

Hypothesis testing

H_0: The average 10 mile run time was the same for 2006 and 2012.

H_A: The average 10 mile run time for 2012 was *different* than that of 2006.

We call H_0 the **null hypothesis** and H_A the **alternative hypothesis**.

Hypothesis testing

Mathematical notation

Let μ_{12} denote the average run time for 2012. Then the two hypothesis are:

H_0: $\mu_{12} = 93.29$

H_A: $\mu_{12} \neq 93.29$

where 93.29 minutes is the average 10 mile time for all runners in the 2006 Cherry Blossom Run.

Testing hypotheses using confidence interval

Cherry Blossom Run

- Null hypothesis $\mu_{12} = 93.29$
- Sample mean $\bar{x}_{12} = 95.61$
- The difference between the $\bar{x}_{12}$ and 93.29 could be due to **sampling variation**
- So we turn our attention to the 95% confidence interval for the time average is

$$(92.45, 98.77)$$

Conclusion: Since 93.29 is in the 95% confidence interval, we can not reject the null hypothesis. We have not found strong evidence that the average running time in 2012 is different from 93.29.

Testing hypotheses using confidence interval

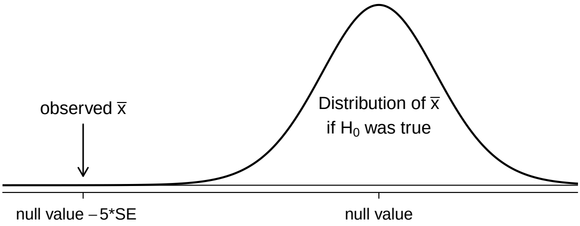
- The null value (the parameter value under the null hypothesis) is in the 95% confidence interval but just barely, so we would not reject H_0. However, we might like to somehow say, quantitatively, that it was a close decision.
- The null value is very far outside of the interval, so we reject H_0. However, we want to communicate that, not only did we reject the null hypothesis, but it wasn't even close.

Significance level- p-value

Heuristics

- We do not reject H_0 unless we have strong evidence.
- What precisely does *strong evidence* mean?
  - As a general rule of thumb, for those cases where the null hypothesis is actually true, we do not want to incorrectly reject H_0 more than $\alpha\%$ of the time.
  - α is called the significance level.

p-value



Hypothesis testing- p-values

Q How long do students sleep?

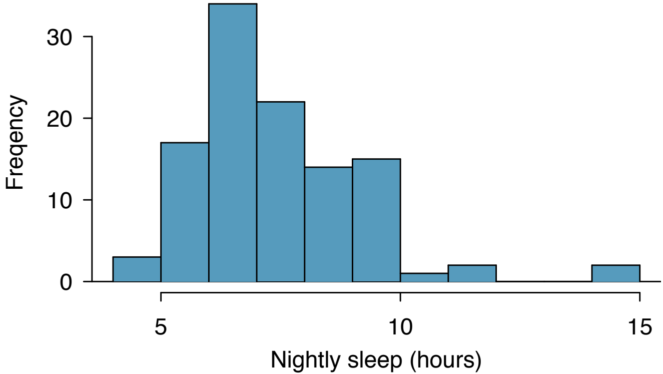


Figure: Distribution of a night of sleep for 110 college students.

Hypothesis testing- p-values

- A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night.
- The average sleep time in the sample is $\bar{x} = 7.42$ hours.
- How can we check is the students average more than 7 hours of sleep?

Hypothesis testing

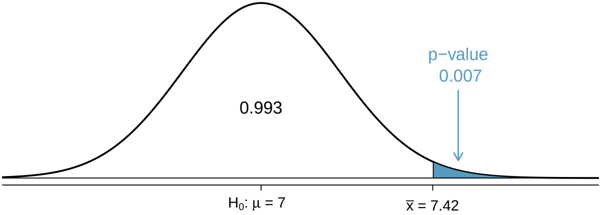
H_0: $\mu = 7.$

H_A: $\mu > 7.$

Evaluate the $Z-$score

$$Z \quad = \quad \frac{\bar{x} - \text{null value}}{SE_{\bar{x}}} = \frac{7.42 - 7}{0.17} = 2.47$$

$$\text{p-value} = P(Z > 2.47) = 0.07$$



Hypothesis testing- p-value

- If the null hypothesis is true, the probability of observing such a large sample mean for a sample of 110 students is only 0.007.
- We evaluate the hypotheses by comparing the p-value to the significance level.
- Since $p - value = 0.007 < 0.05 = \alpha$ we reject the null hypothesis.

The p-value quantifies how strongly the data favour H_A over H_0. A small p-value (usually < 0.05) corresponds to sufficient evidence to reject H_0 in favour of H_A.

Hypothesis testing - p-value

$\alpha-$ **level Hypothesis Testing**

H_0: $\mu = \mu_0$.

$$H_A: \begin{cases} \mu > \mu_0 & \text{(upper-tail alternative)} \\ \mu \neq \mu_0 & \text{(two-tailed alternative)} \\ \mu < \mu_0 & \text{(lower-tail alternative)} \end{cases}$$

Test statistic: $Z = \frac{\hat{x}-\mu_0}{SE}$

We reject H_0 when:

- $P(Z > z) < \alpha$ (upper-tail alternative)
- $P(|Z| > z) < \alpha$ (two-tailed alternative)
- $P(Z < -z) < \alpha$ (lower-tail alternative)

Decision errors

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	okay	Type 1 Error
	H_A true	Type 2 Error	okay

- **Type 1 Error** is rejecting the null hypothesis when H_0 is actually true.
- **Type 2 Error** is failing to reject the null hypothesis when the alternative is actually true.