Multiple random variables

Multiple random variables

We essentially always consider multiple random variables at once.

$\longrightarrow$ The key concepts: Joint, conditional and marginal distributions, and independence of RVs.

Let X and Y be discrete random variables.

$\longrightarrow$ Joint distribution:

$p_{XY}(x,y) = \Pr(X = x \text{ and } Y = y)$

$\longrightarrow$ Marginal distributions:

$p_X(x) = \Pr(X = x) = \sum_y p_{XY}(x,y)$
$p_Y(y) = \Pr(Y = y) = \sum_x p_{XY}(x,y)$

$\longrightarrow$ Conditional distributions:

$p_{X|Y=y}(x) = \Pr(X = x \mid Y = y) = p_{XY}(x,y) \,/\, p_Y(y)$

Example

Sample a couple who are both carriers of some disease gene.

X = number of children they have
Y = number of affected children they have

$p_{XY}(x,y)$		x = 0	1	2	3	4	5	$p_Y(y)$
y	0	0.160	0.248	0.124	0.063	0.025	0.014	0.634
	1	0	0.082	0.082	0.063	0.034	0.024	0.285
	2	0	0	0.014	0.021	0.017	0.016	0.068
	3	0	0	0	0.003	0.004	0.005	0.012
	4	0	0	0	0	0.000	0.001	0.001
	5	0	0	0	0	0	0.000	0.000
	$p_X(x)$	0.160	0.330	0.220	0.150	0.080	0.060	

Pr(Y = y | X = 2)

$p_{XY}(x,y)$		x = 0	1	2	3	4	5	$p_Y(y)$
y	0	0.160	0.248	0.124	0.063	0.025	0.014	0.634
	1	0	0.082	0.082	0.063	0.034	0.024	0.285
	2	0	0	0.014	0.021	0.017	0.016	0.068
	3	0	0	0	0.003	0.004	0.005	0.012
	4	0	0	0	0	0.000	0.001	0.001
	5	0	0	0	0	0	0.000	0.000
	$p_X(x)$	0.160	0.330	0.220	0.150	0.080	0.060	

y	0	1	2	3	4	5
Pr(Y=y | X=2)	0.564	0.373	0.064	0.000	0.000	0.000

$\Pr(X = x \mid Y = 1)$

		x						
$p_{XY}(x,y)$		0	1	2	3	4	5	$p_Y(y)$
	0	0.160	0.248	0.124	0.063	0.025	0.014	0.634
	1	0	0.082	0.082	0.063	0.034	0.024	0.285
y	2	0	0	0.014	0.021	0.017	0.016	0.068
	3	0	0	0	0.003	0.004	0.005	0.012
	4	0	0	0	0	0.000	0.001	0.001
	5	0	0	0	0	0	0.000	0.000
	$p_X(x)$	0.160	0.330	0.220	0.150	0.080	0.060	

x	0	1	2	3	4	5
$\Pr(X=x \mid Y=1)$	0.000	0.288	0.288	0.221	0.119	0.084

Independence

Random variables X and Y are independent if

⟶ $p_{XY}(x,y) = p_X(x)\, p_Y(y)$

for every pair x,y.

In other words/symbols:

⟶ $\Pr(X = x \text{ and } Y = y) = \Pr(X = x)\Pr(Y = y)$

for every pair x,y.

Equivalently,

⟶ $\Pr(X = x \mid Y = y) = \Pr(X = x)$

for all x,y.

Example

Sample a subject from some high-risk population.

X = 1 if the subject is infected with virus A, and = 0 otherwise

Y = 1 if the subject is infected with virus B, and = 0 otherwise

		x = 0	x = 1	
$p_{XY}(x,y)$		0	1	$p_Y(y)$
y	0	0.72	0.18	0.90
	1	0.08	0.02	0.10
	$p_X(x)$	0.80	0.20	

Continuous random variables

Continuous random variables have joint densities, $f_{XY}(x,y)$.

⟶ The marginal densities are obtained by integration:

$$f_X(x) = \int f_{XY}(x, y)\, dy \quad \text{and} \quad f_Y(y) = \int f_{XY}(x, y)\, dx$$

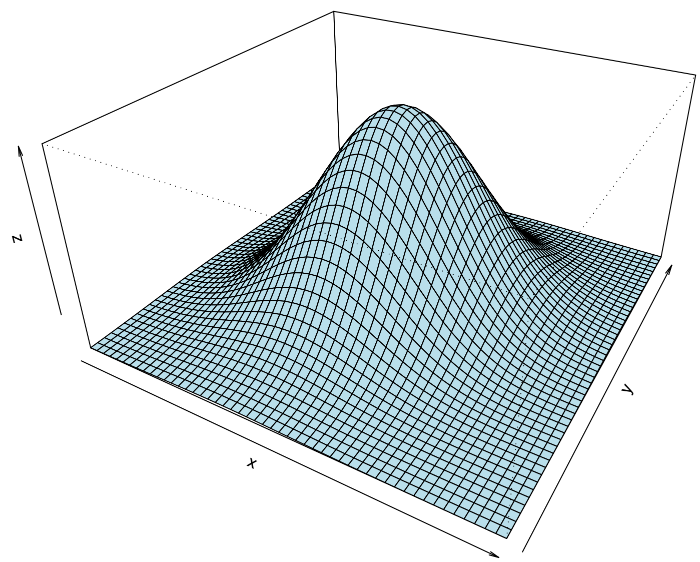
⟶ Conditional density:

$$f_{X|Y=y}(x) = f_{XY}(x, y)/f_Y(y)$$

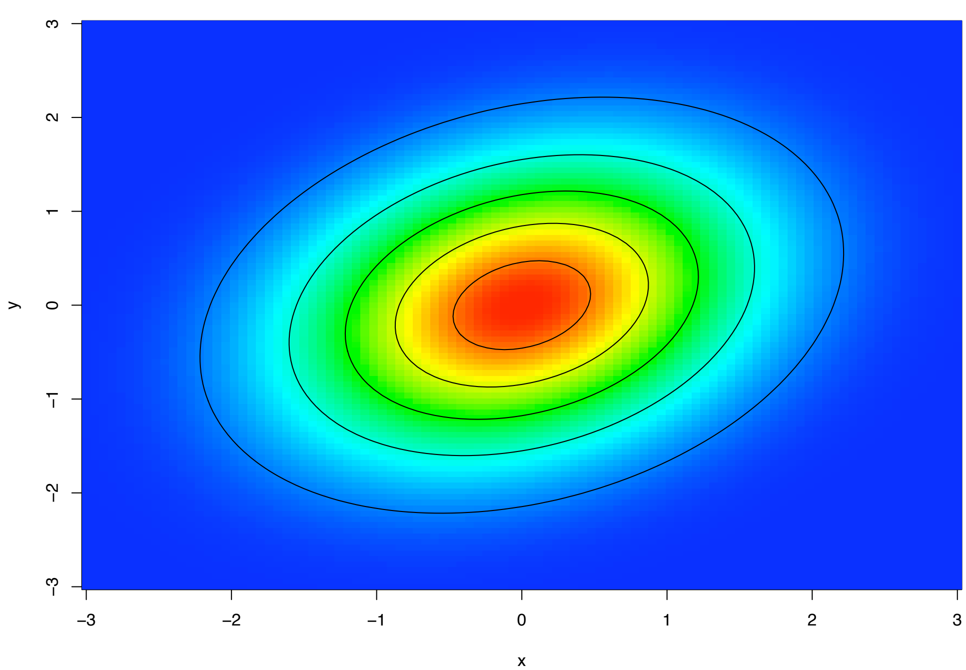
⟶ *X* and *Y* are independent if:

$$f_{XY}(x,y) = f_X(x)\, f_Y(y) \qquad \text{for all } x,y.$$

The bivariate normal distribution



The bivariate normal distribution



IID

More jargon:

Random variables X_1, X_2, X_3, ..., X_n are said to be independent and identically distributed (iid) if

$\longrightarrow$ they are independent,

$\longrightarrow$ they all have the same distribution.

Usually such RVs are generated by

$\longrightarrow$ repeated independent measurements, or

$\longrightarrow$ random sampling from a large population.

Means and SDs

$\longrightarrow$ Mean and SD of sums of random variables:

$E(\sum_i X_i) = \sum_i E(X_i)$ no matter what

$SD(\sum_i X_i) = \sqrt{\sum_i \{SD(X_i)\}^2}$ if the X_i are independent

$\longrightarrow$ Mean and SD of means of random variables:

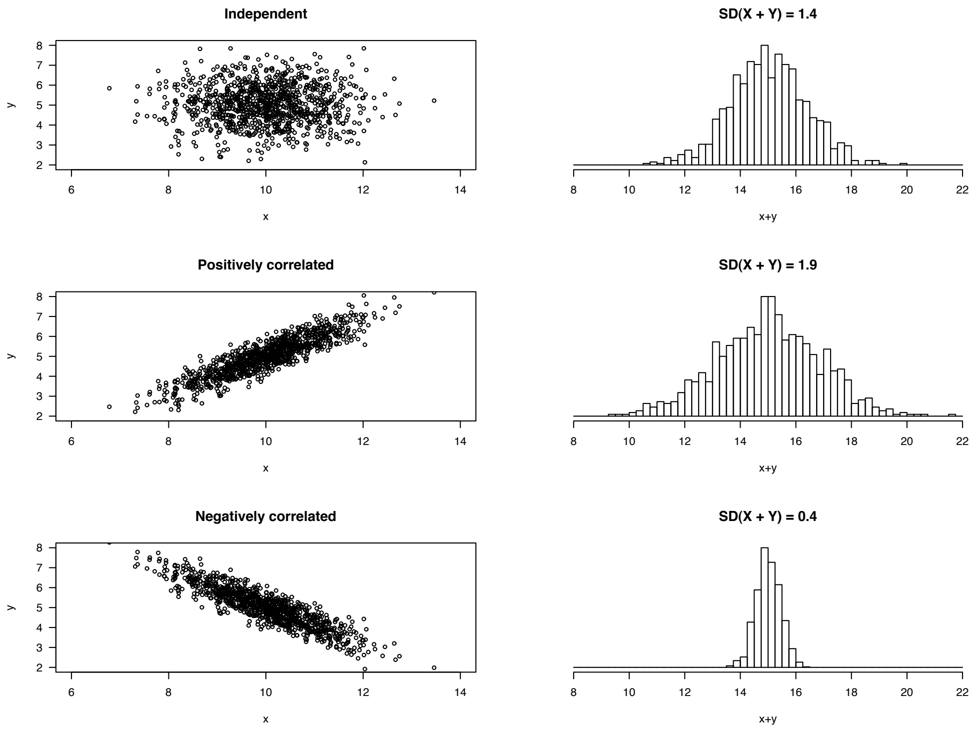
$E(\sum_i X_i \,/\, n) = \sum_i E(X_i)/n$ no matter what

$SD(\sum_i X_i/n) = \sqrt{\sum_i \{SD(X_i)\}^2}/n$ if the X_i are independent

$\longrightarrow$ If the X_i are iid with mean μ and SD σ:

$E(\sum_i X_i \,/\, n) = \mu$ and $SD(\sum_i X_i \,/\, n) = \sigma/\sqrt{n}$

Example



Sampling distributions

Populations and samples

⟶ We are interested in the distribution of measurements in an underlying (possibly hypothetical) population.

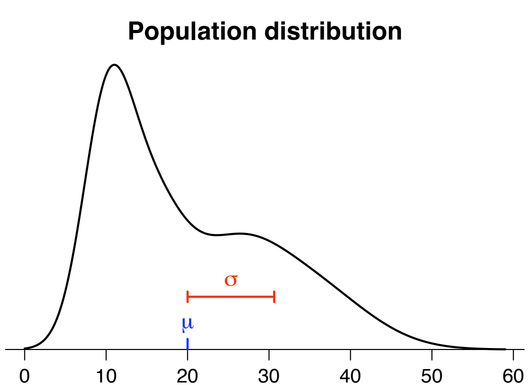
Examples:

- ○ Infinite number of mice from strain A; cytokine response to treatment.
- ○ All T cells in a person; respond or not to an antigen.
- ○ All possible samples from the Baltimore water supply; concentration of cryptospiridium.
- ○ All possible samples of a particular type of cancer tissue; expression of a certain gene.

⟶ We can't see the entire population (whether it is real or hypothetical), but we can see a random sample of the population (perhaps a set of independent, replicated measurements).

Parameters

We are interested in the population distribution or, in particular, certain numerical attributes of the population distribution, called parameters.



⟶ Examples:

- ○ mean
- ○ median
- ○ SD
- ○ proportion = 1
- ○ proportion > 40
- ○ geometric mean
- ○ 95th percentile

Parameters are usually assigned greek letters (like θ, μ, and σ).

Sample data

We make n independent measurements (or draw a random sample of size n). This gives $X_1, X_2, \ldots, X_n$ independent and identically distributed (iid), following the population distribution.

$\longrightarrow$ Statistic:
A numerical summary (function) of the X's. For example, the sample mean, sample SD, etc.

$\longrightarrow$ Estimator:
A statistic, viewed as estimating some population parameter.

We write:

$\overline{X} = \hat{\mu}$ as an estimator of μ, $S = \hat{\sigma}$ as an estimator of σ, $\hat{p}$ as an estimator of p, $\hat{\theta}$ as an estimator of θ, ...

Parameters, estimators, estimates

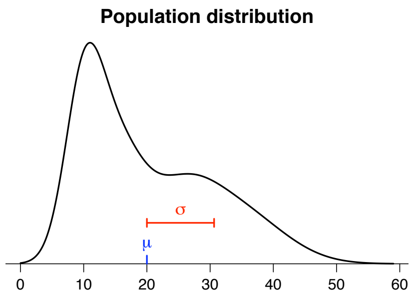
μ
- The population mean
- A parameter
- A fixed quantity
- Unknown, but what we want to know

$\overline{X}$
- The sample mean
- An estimator of μ
- A function of the data (the X's)
- A random quantity

$\overline{x}$
- The observed sample mean
- An estimate of μ
- A particular realization of the estimator, $\overline{X}$
- A fixed quantity, but the result of a random process.

Estimators are random variables

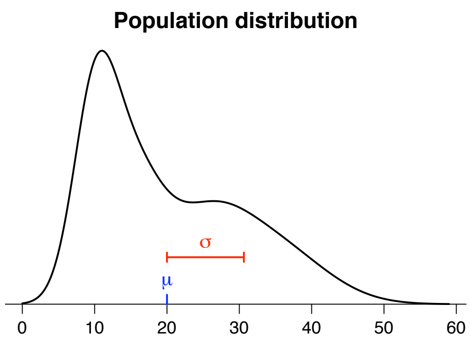
Estimators have distributions, means, SDs, etc.



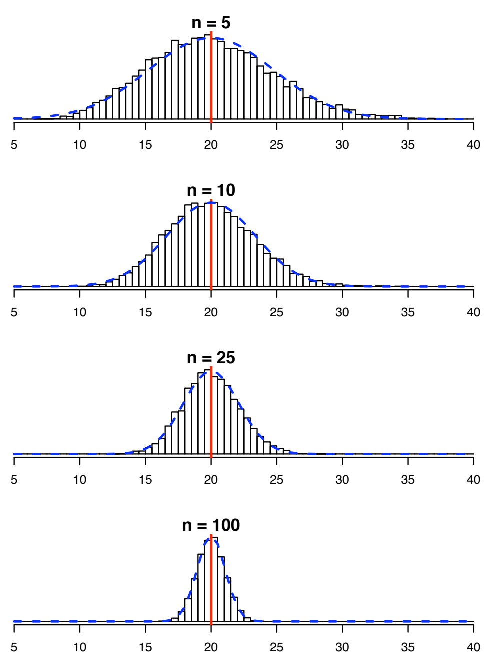
$\longrightarrow \quad X_1, X_2, \ldots, X_{10} \quad \longrightarrow \quad \overline{X}$

3.8	8.0	9.9	13.1	15.5	16.6	22.3	25.4	31.0	40.0	$\longrightarrow$	18.6
6.0	10.6	13.8	17.1	20.2	22.5	22.9	28.6	33.1	36.7	$\longrightarrow$	21.2
8.1	9.0	9.5	12.2	13.3	20.5	20.8	30.3	31.6	34.6	$\longrightarrow$	19.0
4.2	10.3	11.0	13.9	16.5	18.2	18.9	20.4	28.4	34.4	$\longrightarrow$	17.6
8.4	15.2	17.1	17.2	21.2	23.0	26.7	28.2	32.8	38.0	$\longrightarrow$	22.8

Sampling distribution



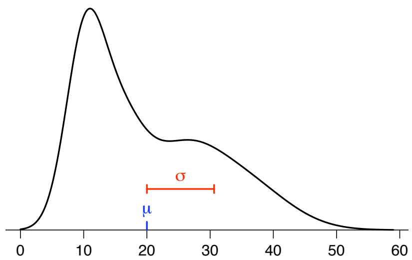
Distribution of $\overline{X}$

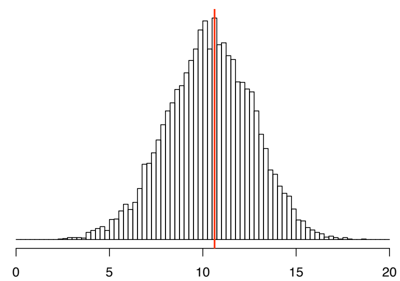


The sampling distribution depends on:

- The type of statistic
- The population distribution
- The sample size

Bias, SE, RMSE





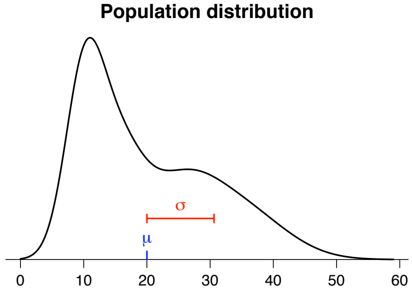
Consider $\hat{\theta}$, an estimator of the parameter θ.

$\longrightarrow$ Bias: $\mathrm{E}(\hat{\theta} - \theta) = \mathrm{E}(\hat{\theta}) - \theta$.

$\longrightarrow$ Standard error (SE): $\mathrm{SE}(\hat{\theta}) = \mathrm{SD}(\hat{\theta})$.

$\longrightarrow$ RMS error (RMSE): $\sqrt{\mathrm{E}\{(\hat{\theta} - \theta)^2\}} = \sqrt{(\text{bias})^2 + (\mathrm{SE})^2}$.

The sample mean



Assume $X_1, X_2, \ldots, X_n$ are iid with mean μ and SD σ.

$\longrightarrow$ Mean of $\overline{X}$ = $\mathrm{E}(\overline{X}) = \mu$.

$\longrightarrow$ Bias = $\mathrm{E}(\overline{X}) - \mu = 0$.

$\longrightarrow$ SE of $\overline{X}$ = $\mathrm{SD}(\overline{X}) = \sigma/\sqrt{n}$.

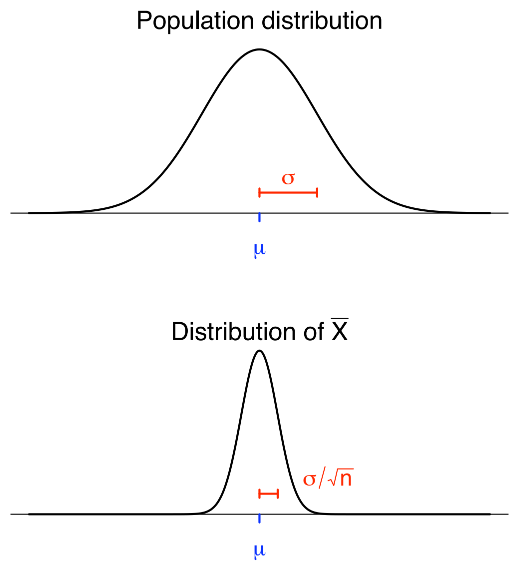
$\longrightarrow$ RMS error of $\overline{X}$:

$\sqrt{(\text{bias})^2 + (\mathrm{SE})^2} = \sigma/\sqrt{n}$.

If the population is normally distributed

If X_1, X_2, ..., X_n are iid Normal(μ,σ), then

$\longrightarrow$ $\overline{X} \sim \text{Normal}(\mu, \sigma/\sqrt{n})$.

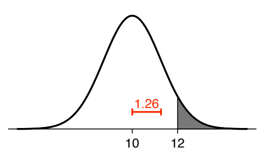


Example

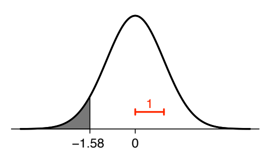
Suppose X_1, X_2, ..., X_{10} are iid Normal(mean=10,SD=4)

Then $\overline{X} \sim$ Normal(mean=10, SD $\approx$ 1.26). Let $Z = (\overline{X} - 10)/1.26$.

Pr($\overline{X} > 12$)?

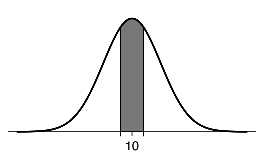


$\approx$

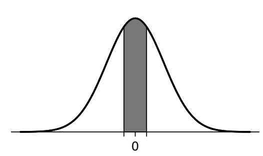


$\approx$ 5.7%

Pr($9.5 < \overline{X} < 10.5$)?

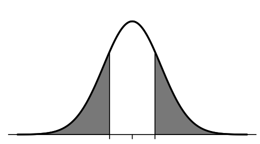


$\approx$

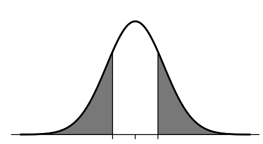


$\approx$ 31%

Pr($|\overline{X} - 10| > 1$)?



$\approx$



$\approx$ 43%

Central limit theorm

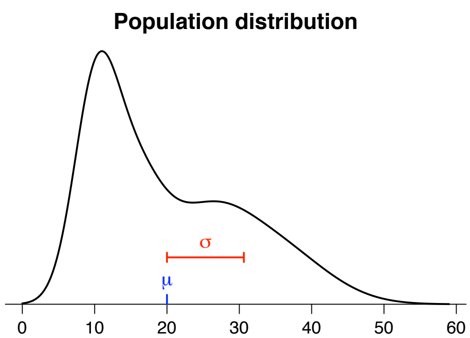
⟶ If $X_1, X_2, \ldots, X_n$ are iid with mean μ and SD σ, and the sample size (n) is large, then

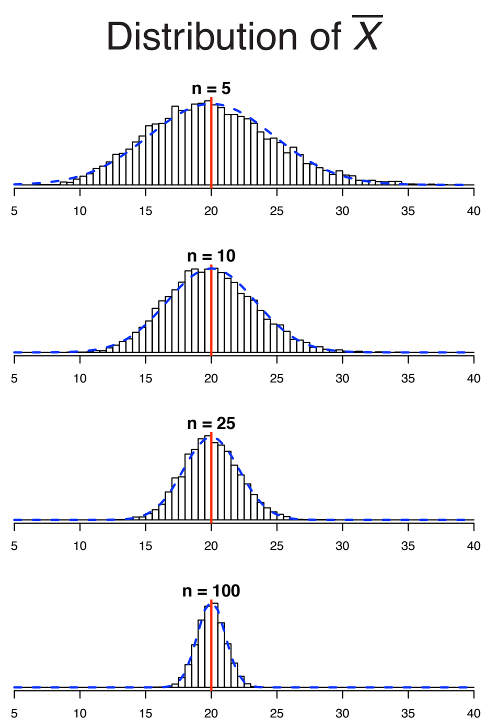
$$\overline{X} \text{ is approximately Normal}(\mu, \sigma/\sqrt{n}).$$

⟶ How large is large?

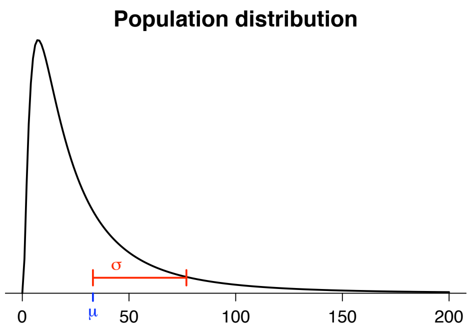
It depends on the population distribution.
(But, generally, not too large.)

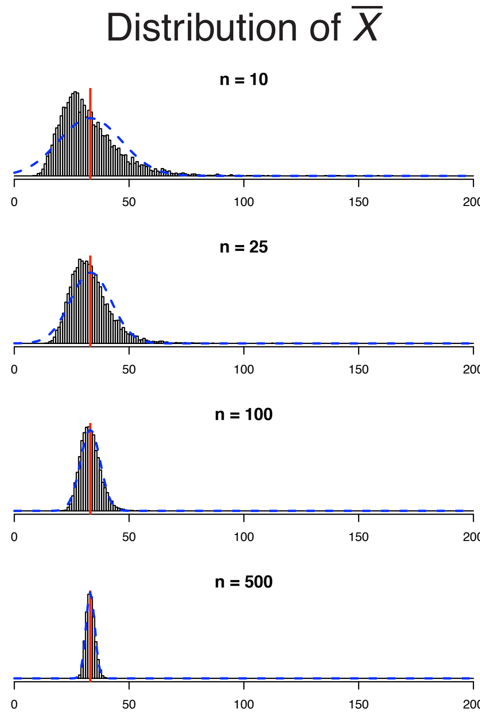
Example 1



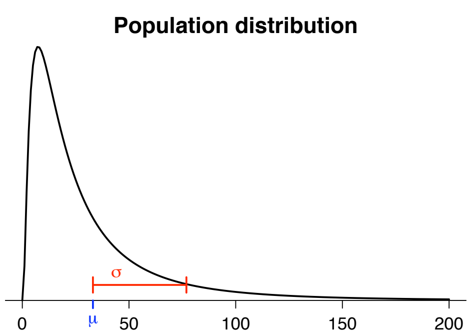


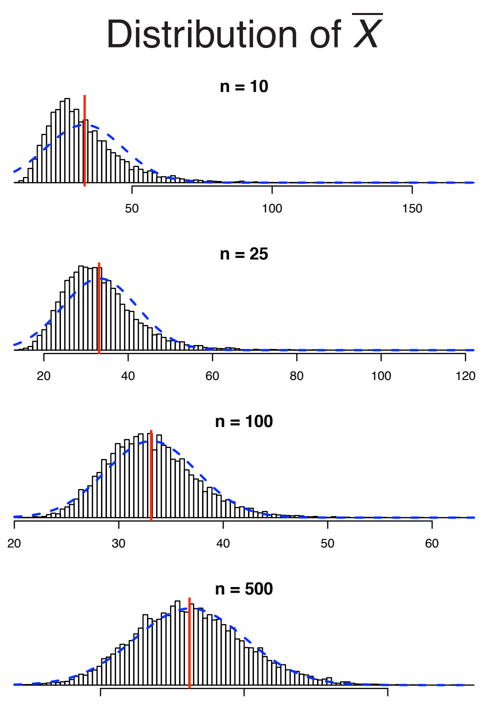
Example 2



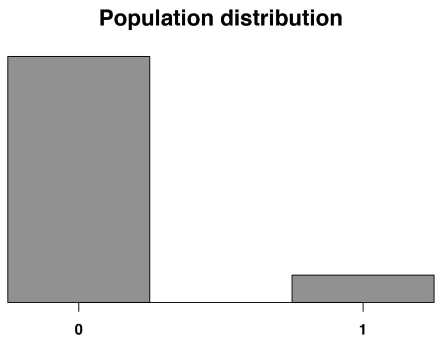


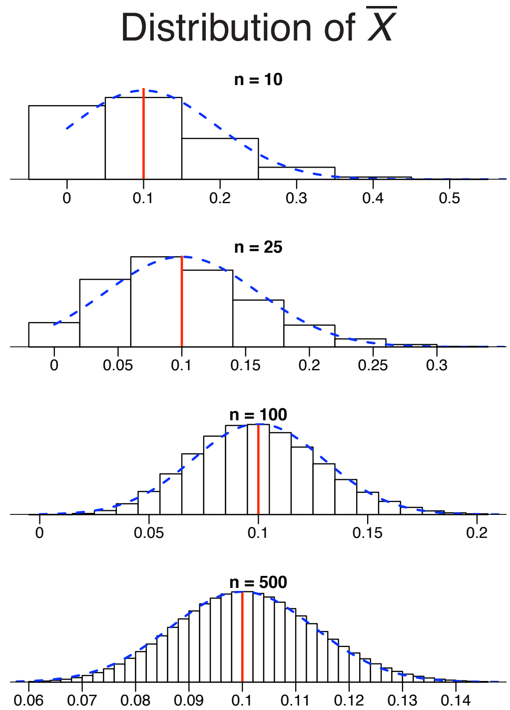
Example 2 (rescaled)





Example 3





$\{X_i\}$ iid

$\Pr(X_i = 0) = 90\%$
$\Pr(X_i = 1) = 10\%$

$E(X_i) = 0.1$; $SD(X_i) = 0.3$

$\sum X_i \sim \text{Binomial(n, p)}$

$\rightarrow$ $\overline{X}$ = proportion of 1's

The sample SD

$\longrightarrow$ Why use (n – 1) in the sample SD?

$$S = \sqrt{\frac{\sum(X_i - \overline{X})^2}{n-1}}$$

$\longrightarrow$ If $\{X_i\}$ are iid with mean μ and SD σ, then

- $E(S^2) = \sigma^2$
- $E\{\,\frac{n-1}{n}\,S^2\,\} = \frac{n-1}{n}\,\sigma^2 < \sigma^2$

$\longrightarrow$ In other words:

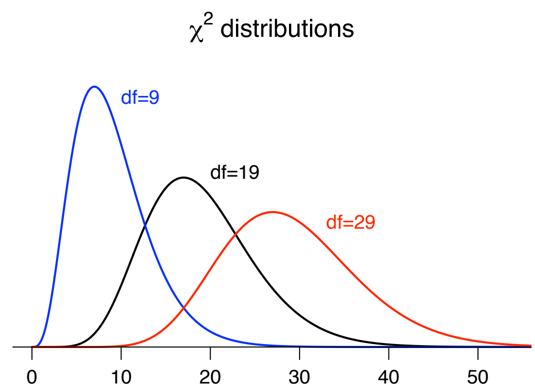
- $\text{Bias}(S^2) = 0$
- $\text{Bias}\left(\,\frac{n-1}{n}\,S^2\,\right) = \frac{n-1}{n}\,\sigma^2 - \sigma^2 = -\frac{1}{n}\,\sigma^2$

The distribution of the sample SD

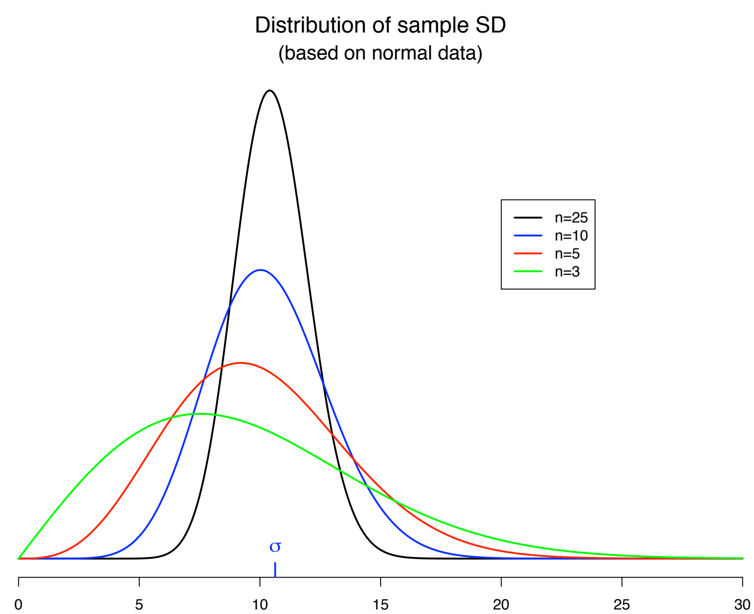
⟶ If $X_1, X_2, \ldots, X_n$ are iid Normal(μ, σ), then the sample SD S satisfies

$$(n-1)\, S^2/\sigma^2 \sim \chi^2_{n-1}$$

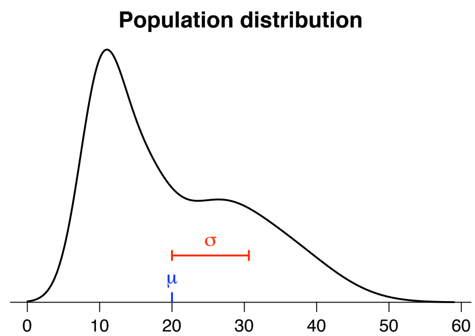
(When the X_i are not normally distributed, this is not true.)

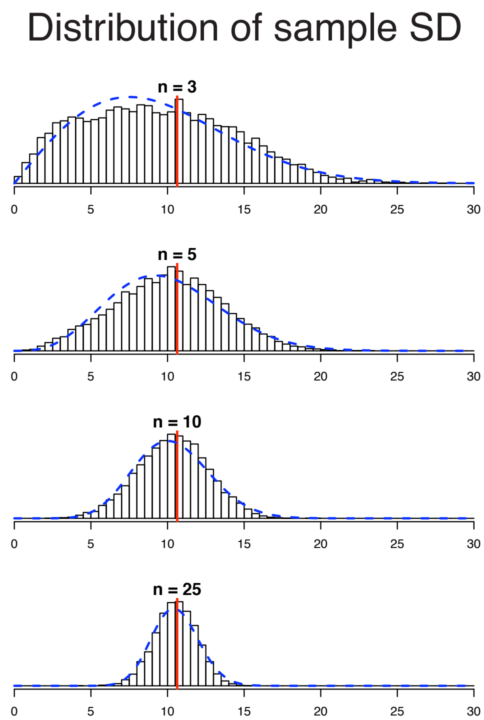


Example



A non-normal example





Inference about one group

Review

$\longrightarrow$ If $X_1, \ldots, X_n$ have mean μ and SD σ, then

$E(\overline{X}) = \mu$ no matter what

$SD(\overline{X}) = \sigma/\sqrt{n}$ if the X's are independent

$\longrightarrow$ If $X_1, \ldots, X_n$ are iid Normal(mean=μ, SD=σ), then

$$\overline{X} \sim \text{Normal}(\text{mean} = \mu, \text{SD} = \sigma/\sqrt{n}).$$

$\longrightarrow$ If $X_1, \ldots, X_n$ are iid with mean μ and SD σ and the sample size n is large, then

$$\overline{X} \sim \text{Normal}(\text{mean} = \mu, \text{SD} = \sigma/\sqrt{n}).$$

Confidence intervals

Suppose we measure some response in 100 male subjects, and find that the sample average ($\bar{x}$) is 3.52 and sample SD (s) is 1.61.

Our estimate of the SE of the sample mean is $1.61/\sqrt{100} = 0.161$.

A 95% confidence interval for the population mean (μ) is roughly

$$3.52 \pm (2 \times 0.16) = 3.52 \pm 0.32 = (3.20, 3.84).$$

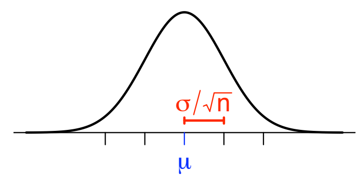
What does this mean?

Confidence intervals

Suppose that $X_1, \ldots, X_n$ are iid Normal(mean=μ, SD=σ).
Suppose that we actually know σ.

Then $\bar{X} \sim$ Normal(mean=μ, SD=$\sigma/\sqrt{n}$) $\quad$ σ is known but μ is not!

$\longrightarrow$ How close is $\overline{X}$ to μ?



$$\Pr\left(\frac{|\overline{X}-\mu|}{\sigma/\sqrt{n}} \le 1.96\right) = 95\%$$

$$\Pr\left(\frac{-1.96\,\sigma}{\sqrt{n}} \le \overline{X}-\mu \le \frac{1.96\,\sigma}{\sqrt{n}}\right) = 95\%$$

$$\Pr\left(\overline{X}-\frac{1.96\,\sigma}{\sqrt{n}} \le \mu \le \overline{X}+\frac{1.96\,\sigma}{\sqrt{n}}\right) = 95\%$$

What is a confidence interval?

A 95% confidence interval is an interval calculated from the data that in advance has a 95% chance of covering the population parameter.

In advance, $\overline{X} \pm 1.96\sigma/\sqrt{n}$ has a 95% chance of covering μ.

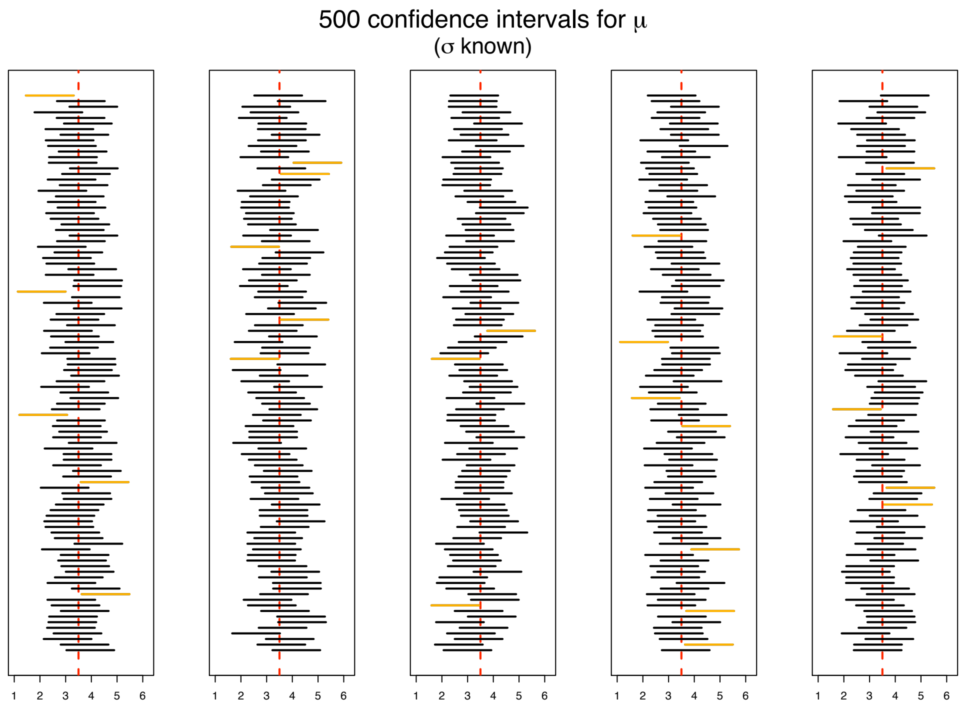
Thus, it is called a 95% confidence interval for μ.

Note that, after the data is gathered (for instance, n=100, $\bar{x} = 3.52$, $\sigma = 1.61$), the interval becomes fixed:

$$\bar{x} \pm 1.96\sigma/\sqrt{n} = 3.52 \pm 0.32.$$

We can't say that there's a 95% chance that μ is in the interval 3.52 ± 0.32. It either is or it isn't; we just don't know.

What is a confidence interval?



Longer and shorter intervals

⟶ If we use 1.64 in place of 1.96, we get shorter intervals with lower confidence.

Since $\Pr\left(\frac{|\overline{X}-\mu|}{\sigma/\sqrt{n}} \leq 1.64\right) = 90\%$,

$\overline{X} \pm 1.64\sigma/\sqrt{n}$ is a 90% confidence interval for μ.

⟶ If we use 2.58 in place of 1.96, we get longer intervals with higher confidence.

Since $\Pr\left(\frac{|\overline{X}-\mu|}{\sigma/\sqrt{n}} \leq 2.58\right) = 99\%$,

$\overline{X} \pm 2.58\sigma/\sqrt{n}$ is a 99% confidence interval for μ.

What is a confidence interval? (cont)

A 95% confidence interval is obtained from a procedure for producing an interval, based on data, that 95% of the time will produce an interval covering the population parameter.

In advance, there's a 95% chance that the interval will cover the population parameter.

After the data has been collected, the confidence interval either contains the parameter or it doesn't.

Thus we talk about confidence rather than probability.

But we don't know the SD

Use of $\overline{X} \pm 1.96\,\sigma/\sqrt{n}$ as a 95% confidence interval for μ requires knowledge of σ.

That the above is a 95% confidence interval for μ is a result of the following:

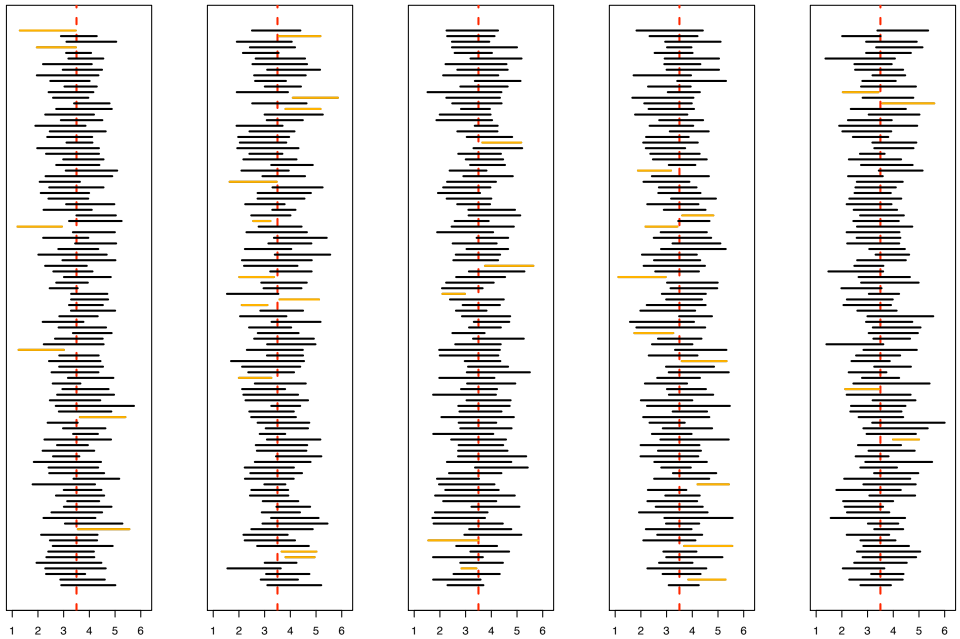
$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal(0,1)}$$

What if we don't know σ?

$\longrightarrow$ We plug in the sample SD S, but then we need to widen the intervals to account for the uncertainty in S.

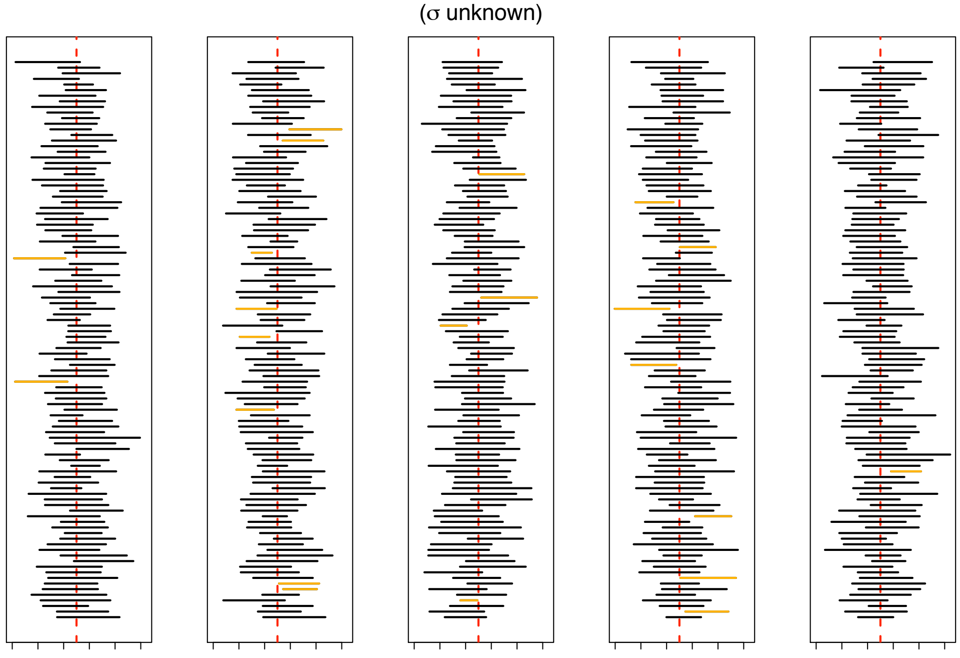
What is a confidence interval? (cont)

500 BAD confidence intervals for μ
(σ unknown)



What is a confidence interval? (cont)

500 confidence intervals for μ
(σ unknown)

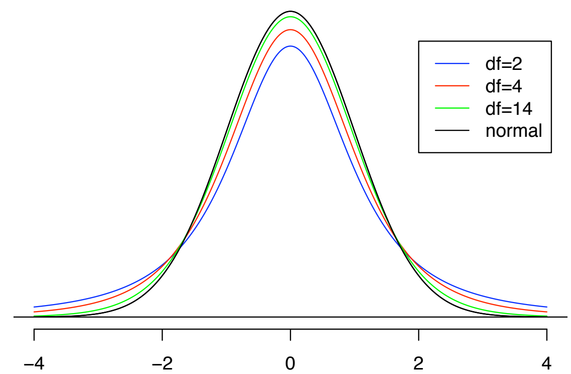


The Student t distribution

If $X_1, X_2, \ldots X_n$ are iid Normal(mean=μ, SD=σ), then

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(df = n - 1)$$

Discovered by William Gossett ("Student") who worked for Guinness.



⟶ `qt(0.975,9)` returns `2.26` (compare to 1.96)

⟶ `pt(1.96,9)-pt(-1.96,9)` returns `0.918` (compare to 0.95)

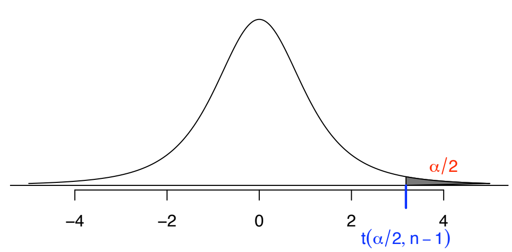
The t interval

If $X_1, \ldots, X_n$ are iid Normal(mean=μ, SD=σ), then

$$\overline{X} \pm t(\alpha/2, n-1)\ S/\sqrt{n}$$

is a $1 - \alpha$ confidence interval for μ.

⟶ $t(\alpha/2, n-1)$ is the $1 - \alpha/2$ quantile of the t distribution with $n - 1$ "degrees of freedom."



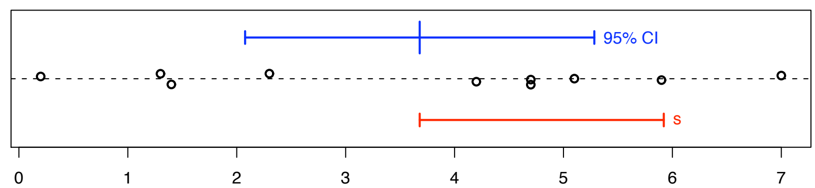
Example 1

Suppose we have measured some response in 10 male subjects, and obtained the following numbers:

Data		
0.2 1.3 1.4 2.3 4.2	$\bar{x}$ = 3.68	n = 10
4.7 4.7 5.1 5.9 7.0	s = 2.24	`qt(0.975,9)` = 2.26

$\longrightarrow$ 95% confidence interval for μ (the population mean):

$$3.68 \pm 2.26 \times 2.24 \,/\, \sqrt{10} \approx 3.68 \pm 1.60 = (2.1, 5.3)$$



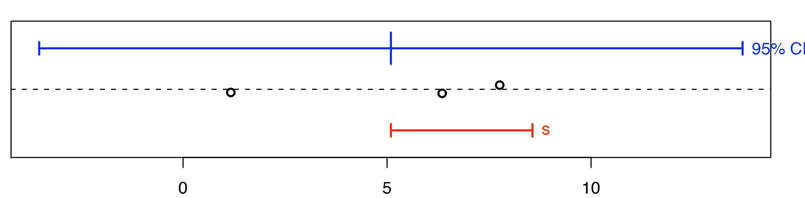
Example 2

Suppose we have measured (by RT-PCR) the $\log_{10}$ expression of a gene in 3 tissue samples, and obtained the following numbers:

Data		
1.17 6.35 7.76	$\bar{x}$ = 5.09	n = 3
	s = 3.47	`qt(0.975,2)` = 4.30

$\longrightarrow$ 95% confidence interval for μ (the population mean):

$$5.09 \pm 4.30 \times 3.47 \,/\, \sqrt{3} \approx 5.09 \pm 8.62 = (-3.5, 13.7)$$



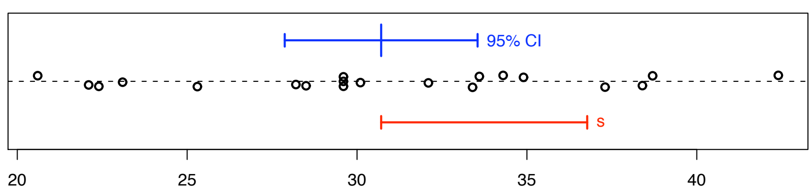
Example 3

Suppose we have weighed the mass of tumor in 20 mice, and obtained the following numbers

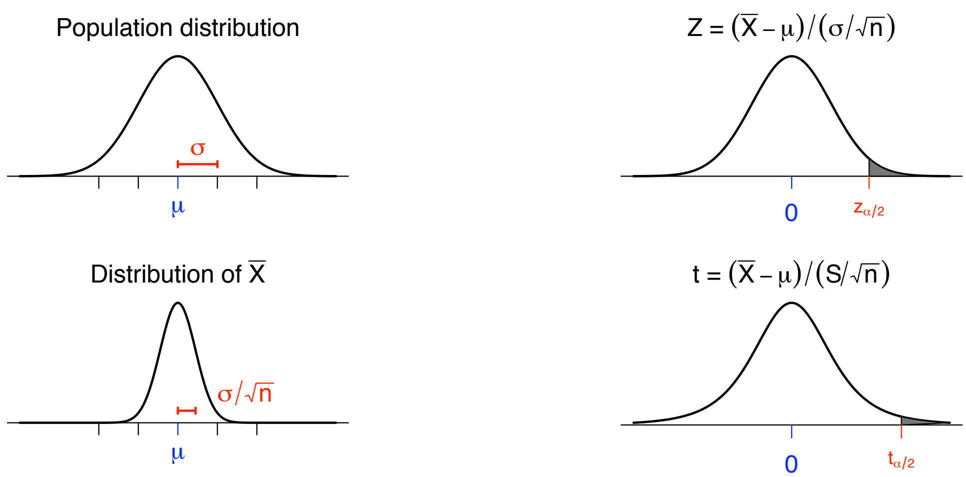
Data						
34.9	28.5	34.3	38.4	29.6	$\bar{x} = 30.7$	$n = 20$
28.2	25.3	...	...	32.1	$s = 6.06$	`qt(0.975,19)` = 2.09

$\longrightarrow$ 95% confidence interval for μ (the population mean):

$$30.7 \pm 2.09 \times 6.06 \,/\, \sqrt{20} \approx 30.7 \pm 2.84 = (27.9, 33.5)$$



Confidence interval for the mean



$X_1, X_2, \ldots, X_n$ independent Normal(μ, σ).

95% confidence interval for μ:

$\overline{X} \pm t\, S/\sqrt{n}$ where t = 97.5 percentile of t distribution with (n – 1) d.f.

Confidence interval for the population SD

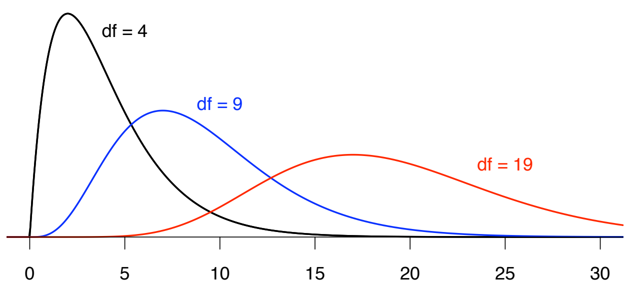
Suppose we observe $X_1, X_2, \ldots, X_n$ iid Normal(μ, σ).

Suppose we wish to create a 95% CI for the population SD, σ.

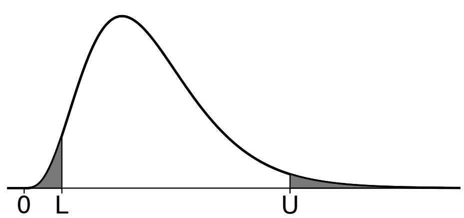
Our estimate of σ is the sample SD, S.

The sampling distribution of S is such that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(\mathrm{df} = n-1)$$



Confidence interval for the population SD



Choose L and U such that

$\Pr\left(L \le \frac{(n-1)S^2}{\sigma^2} \le U\right) = 95\%.$

$$\Pr\left(\frac{1}{U} \le \frac{\sigma^2}{(n-1)S^2} \le \frac{1}{L}\right) = 95\%.$$

$$\Pr\left(\frac{(n-1)S^2}{U} \le \sigma^2 \le \frac{(n-1)S^2}{L}\right) = 95\%.$$

$$\Pr\left(S\sqrt{\frac{n-1}{U}} \le \sigma \le S\sqrt{\frac{n-1}{L}}\right) = 95\%.$$

$\longrightarrow$ $\left(S\sqrt{\frac{n-1}{U}}\,,\ S\sqrt{\frac{n-1}{L}}\right)$ is a 95% CI for σ.

Example

Population A: $n = 10$; sample SD: $s_A = 7.64$

$L =$ `qchisq(0.025,9)` $= 2.70$
$U =$ `qchisq(0.975,9)` $= 19.0$

$\longrightarrow$ 95% CI for σ_A:

$$(7.64 \times \sqrt{\tfrac{9}{19.0}},\ 7.64 \times \sqrt{\tfrac{9}{2.70}}) = (7.64 \times 0.69,\ 7.64 \times 1.83) = (5.3,\ 14.0)$$

Population B: $n = 16$; sample SD: $s_B = 18.1$

$L =$ `qchisq(0.025,15)` $= 6.25$
$U =$ `qchisq(0.975,15)` $= 27.5$

$\longrightarrow$ 95% CI for σ_B:

$$(18.1 \times \sqrt{\tfrac{15}{27.5}},\ 18.1 \times \sqrt{\tfrac{15}{6.25}}) = (18.1 \times 0.74,\ 18.1 \times 1.55) = (13.4,\ 28.1)$$

Tests of hypotheses

Confidence interval: Form an interval (on the basis of data) of plausible values for a population parameter.

Test of hypothesis: Answer a yes or no question regarding a population parameter.

Examples:

$\longrightarrow$ Do the two strains have the same average response?

$\longrightarrow$ Is the concentration of substance X in the water supply above the safe limit?

$\longrightarrow$ Does the treatment have an effect?

Example

We have a quantitative assay for the concentration of antibodies against a certain virus in blood from a mouse.

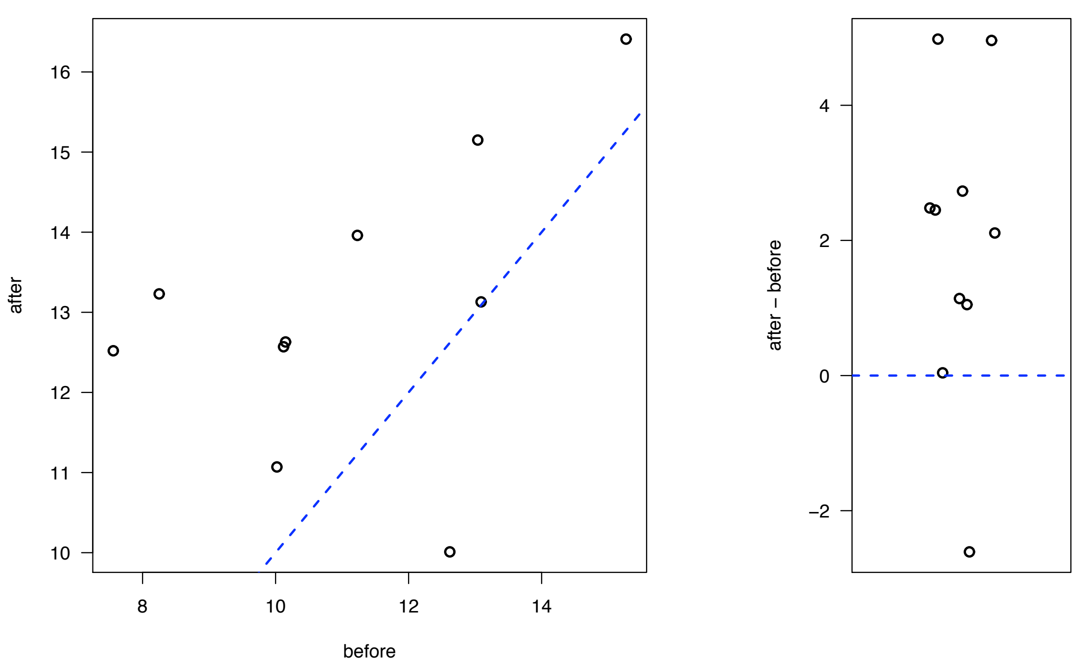
We apply our assay to a set of ten mice before and after the injection of a vaccine. (This is called a “paired” experiment.)

Let X_i denote the differences between the measurements (“after” minus “before”) for mouse i.

We imagine that the X_i are independent and identically distributed Normal(μ, σ).

$\longrightarrow$ Does the vaccine have an effect? In other words: Is $\mu \neq 0$?

The data



Hypothesis testing

We consider two hypotheses:

Null hypothesis, H_0: $\mu = 0$ Alternative hypothesis, H_a: $\mu \neq 0$

Type I error:	Reject H_0 when it is true	(false positive)
Type II error:	Fail to reject H_0 when it is false	(false negative)

We set things up so that a Type I error is a worse error (and so that we are seeking to prove the alternative hypothesis). We want to control the rate (the significance level, α) of such errors.

$\longrightarrow$ Test statistic: $T = (\overline{X} - 0)/(S/\sqrt{10})$

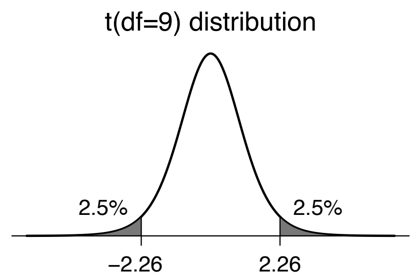
$\longrightarrow$ We reject H_0 if $|T| > t^\star$, where $t^\star$ is chosen so that $\Pr(\text{Reject } H_0 \mid H_0 \text{ is true}) = \Pr(|T| > t^\star \mid \mu = 0) = \alpha$.
(generally α = 5%)

Example (continued)

Under H_0 (i.e., when μ = 0),

$T = (\overline{X} - 0)/(S/\sqrt{10}) \sim t(df = 9)$

We reject H_0 if $|T| > 2.26$.



As a result, if H_0 is true, there's a 5% chance that you'll reject it!

For the observed data:

$\bar{x}$ = 1.93, s = 2.24, n = 10 $\quad$ T = (1.93 - 0) / (2.24/$\sqrt{10}$) = 2.72

$\longrightarrow$ Thus we reject H_0.

The goal

$\longrightarrow$ We seek to prove the alternative hypothesis.

$\longrightarrow$ We are happy if we reject H_0.

$\longrightarrow$ In the case that we reject H_0, we might say:

Either H_0 is false, or a rare event occurred.

Another example

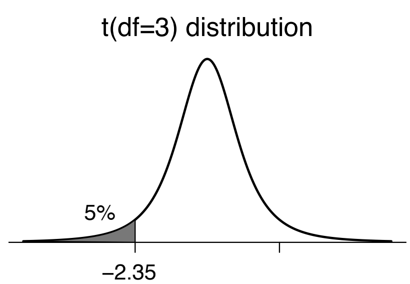
Question: is the concentration of substance X in the water supply above the safe level?

$X_1, X_2, \ldots, X_4 \sim$ iid Normal(μ, σ).

$\longrightarrow$ We want to test H_0: $\mu \geq 6$ (unsafe) versus H_a: $\mu < 6$ (safe).

Test statistic: $T = \dfrac{\overline{X} - 6}{S/\sqrt{4}}$

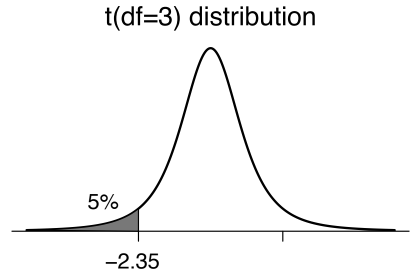
If we wish to have the significance level α = 5%, the rejection region is $T < t^\star = -2.35$.



One-tailed vs two-tailed tests

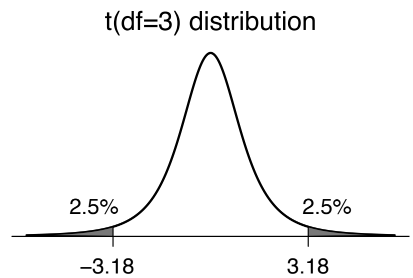
If you are trying to prove that a treatment improves things, you want a one-tailed (or one-sided) test.

You'll reject H_0 only if $T < t^{\star}$.



If you are just looking for a difference, use a two-tailed (or two-sided) test.

You'll reject H_0 if $T < t^{\star}$ or $T > t^{\star}$.



P-values

P-value: $\longrightarrow$ the smallest significance level (α) for which you would fail to reject H_0 with the observed data.

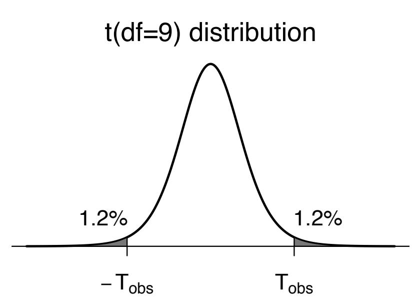
$\longrightarrow$ the probability, if H_0 was true, of receiving data as extreme as what was observed.

$X_1, \ldots, X_{10} \sim$ iid Normal(μ, σ), $\quad$ H_0: $\mu = 0$; H_a: $\mu \neq 0$.

$\bar{x} = 1.93$; $s = 2.24$

$T_{obs} = \frac{1.93-0}{2.24/\sqrt{10}} = 2.72$

P-value = $\Pr(|T| > T_{obs}) = 2.4\%$.



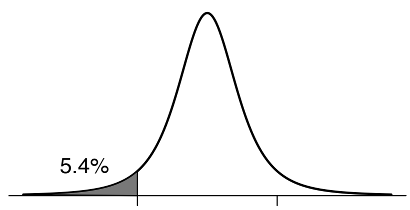
Another example

$X_1, \dots, X_4 \sim \text{Normal}(\mu, \sigma)$ H$_0$: $\mu \geq 6$; H$_a$: $\mu < 6$.

$\bar{x} = 5.51$; s = 0.43

$T_{obs} = \frac{5.51 - 6}{0.43/\sqrt{4}} = -2.28$

P-value = $\Pr(T < T_{obs} \mid \mu = 6) = 5.4\%$.



Recall: We want to prove the alternative hypothesis (i.e., reject H$_0$, receive a small P-value)

Hypothesis tests and confidence intervals

$\longrightarrow$ The 95% confidence interval for μ is the set of values, μ_0, such that the null hypothesis H$_0$: $\mu = \mu_0$ would not be rejected by a two-sided test with $\alpha = 5\%$.

The 95% CI for μ is the set of plausible values of μ. If a value of μ is plausible, then as a null hypothesis, it would not be rejected.

For example:

9.98 9.87 10.05 10.08 9.99 9.90 assumed to be iid Normal(μ,σ)

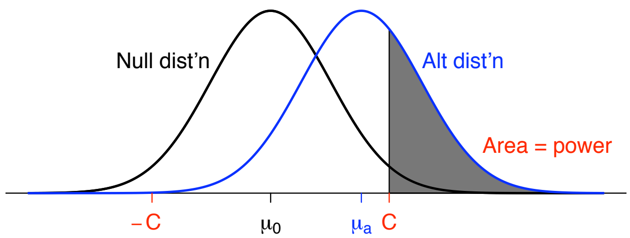
$\bar{x} = 9.98$; s = 0.082; n = 6; `qt(0.975,5)` = 2.57

The 95% CI for μ is

$$9.98 \pm 2.57 \times 0.082 / \sqrt{6} = 9.98 \pm 0.086 = (9.89, 10.06)$$

Power

The power of a test = Pr(reject H_0 | H_0 is false).



The power depends on:
- The null hypothesis and test statistic
- The sample size
- The true value of μ
- The true value of σ

Why "fail to reject"?

If the data are insufficient to reject H_0, we say,

The data are insufficient to reject H_0.

We shouldn't say, *We have proven H_0.*

$\longrightarrow$ We may only have low power to detect anything but extreme differences.

$\longrightarrow$ We control the rate of type I errors ("false positives") at 5% (or whatever), but we may have little or no control over the rate of type II errors.

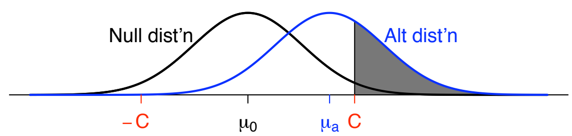
The effect of sample size

Let $X_1, \ldots, X_n$ be iid Normal(μ, σ).

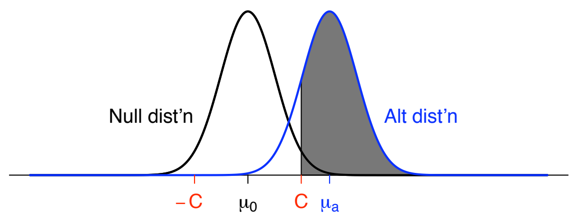
We wish to test $H_0 : \mu = \mu_0$ vs $H_a : \mu \neq \mu_0$.

Imagine $\mu = \mu_a$.

n = 4



n = 16



Test for a proportion

Suppose X $\sim$ Binomial(n, p).

Test $H_0 : p = \frac{1}{2}$ vs $H_a : p \neq \frac{1}{2}$.

Reject H_0 if $X \geq H$ or $X \leq L$.

Choose H and L such that

$$\Pr(X \geq H \mid p = \tfrac{1}{2}) \leq \alpha/2 \quad \text{and} \quad \Pr(X \leq L \mid p = \tfrac{1}{2}) \leq \alpha/2.$$

Thus Pr(Reject H_0 | H_0 is true) $\leq \alpha$.

$\longrightarrow$ The difficulty: The Binomial distribution is hard to work with. Because of its discrete nature, you can't get exactly your desired significance level (α).

Rejection region

Consider X $\sim$ Binomial(n=29, p).

Test of $H_0 : p = \frac{1}{2}$ vs $H_a : p \neq \frac{1}{2}$ at significance level α = 0.05.

Lower critical value:

$\Pr(X \leq 8) = 0.012$

$\Pr(X \leq 9) = 0.031 \;\longrightarrow\; L = 8$

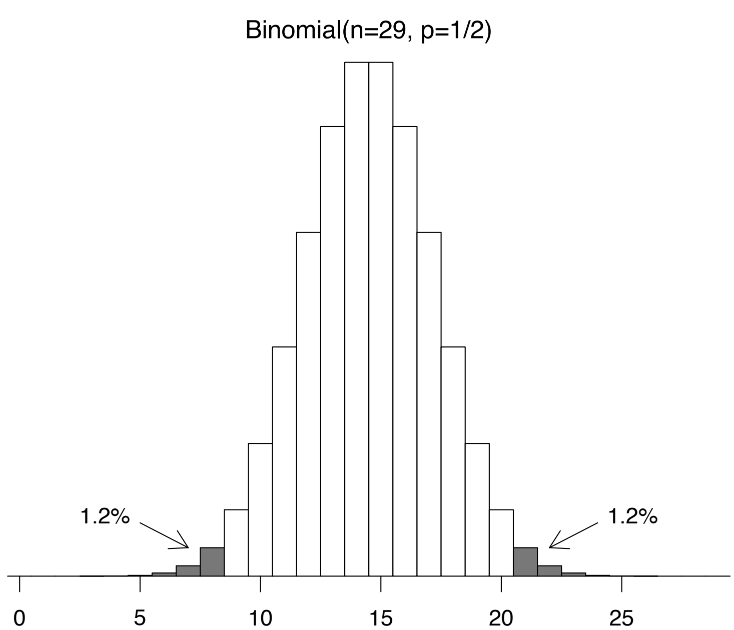
Upper critical value:

$\Pr(X \geq 21) = 0.012$

$\Pr(X \geq 20) = 0.031 \;\longrightarrow\; H = 21$

Reject H_0 if $X \leq 8$ or $X \geq 21$. (For testing $H_0 : p = \frac{1}{2}$, H = n – L)

Binomial(n=29, p=1/2)



Significance level

Consider X $\sim$ Binomial(n=29, p).

Test of $H_0 : p = \frac{1}{2}$ vs $H_a : p \neq \frac{1}{2}$ at significance level $\alpha = 0.05$.

Reject H_0 if $X \leq 8$ or $X \geq 21$.

Actual significance level:

$$\begin{aligned}\alpha &= \Pr(X \leq 8 \text{ or } X \geq 21 \mid p = \tfrac{1}{2}) \\ &= \Pr(X \leq 8 \mid p = \tfrac{1}{2}) + [1 - \Pr(X \leq 20 \mid p = \tfrac{1}{2})] \\ &= 0.024\end{aligned}$$

If we used instead *"Reject H_0 if $X \leq 9$ or $X \geq 20$"*, the significance level would be 0.061!

Confidence interval for a proportion

Suppose X $\sim$ Binomial(n=29, p) and we observe X = 24.

Consider the test of $H_0 : p = p_0$ vs $H_a : p \neq p_0$.

We reject H_0 if

$$\Pr(X \leq 24 \mid p = p_0) \leq \alpha/2 \qquad \text{or} \qquad \Pr(X \geq 24 \mid p = p_0) \leq \alpha/2$$

95% confidence interval for p:

$\longrightarrow$ The set of p_0 for which a two-tailed test of $H_0 : p = p_0$ would not be rejected, for the observed data, with $\alpha = 0.05$.

$\longrightarrow$ The "plausible" values of p.

Example 1

$X \sim$ Binomial(n=29, p); observe X = 24.

Lower bound of 95% confidence interval:

Largest p_0 such that $\Pr(X \geq 24 \mid p = p_0) \leq 0.025$

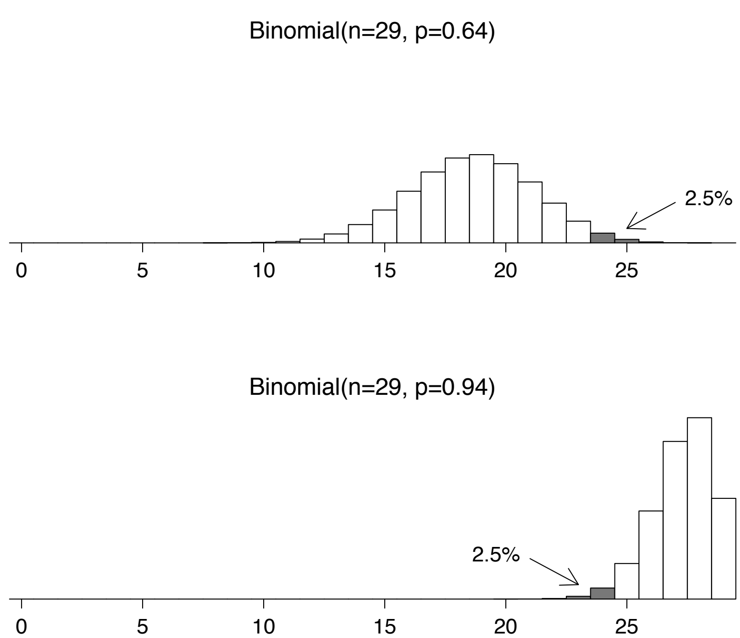
Upper bound of 95% confidence interval:

Smallest p_0 such that $\Pr(X \leq 24 \mid p = p_0) \leq 0.025$

$\longrightarrow$ 95% CI for p: (0.642, 0.942)

Note: $\hat{p}$ = 24/29 = 0.83 is not the midpoint of the CI.

Example 1



Example 2

X $\sim$ Binomial(n=25, p); observe X = 17.

Lower bound of 95% confidence interval:

p_L such that 17 is the 97.5 percentile of Binomial(n=25, p_L)

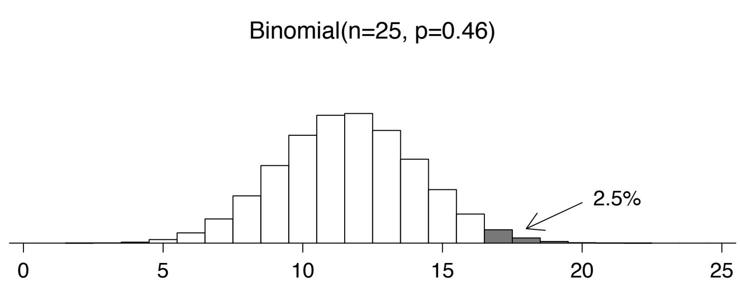
Upper bound of 95% confidence interval:

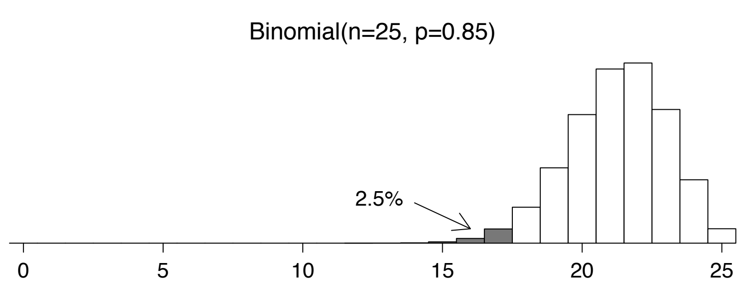
p_H such that 17 is the 2.5 percentile of Binomial(n=25, p_H)

$\longrightarrow$ 95% CI for p: (0.465, 0.851)

Again, $\hat{p}$ = 17/25 = 0.68 is not the midpoint of the CI

Example 2





The case X = 0

Suppose $X \sim \text{Binomial}(n, p)$ and we observe $X = 0$.

Lower limit of 95% confidence interval for p: $\to 0$

Upper limit of 95% confidence interval for p:

p_H such that

$$
\begin{aligned}
& \Pr(X \le 0 \mid p = p_H) = 0.025 \\
\Longrightarrow\ & \Pr(X = 0 \mid p = p_H) = 0.025 \\
\Longrightarrow\ & (1 - p_H)^n = 0.025 \\
\Longrightarrow\ & 1 - p_H = \sqrt[n]{0.025} \\
\Longrightarrow\ & p_H = 1 - \sqrt[n]{0.025}
\end{aligned}
$$

In the case $n = 10$ and $X = 0$, the 95% CI for p is (0, 0.31).

A mad cow example

New York Times, Feb 3, 2004:

> The department [of Agriculture] has not changed last year's plans to test 40,000 cows nationwide this year, out of 30 million slaughtered. Janet Riley, a spokeswoman for the American Meat Institute, which represents slaughterhouses, called that "plenty sufficient from a statistical standpoint."

Suppose that the 40,000 cows tested are chosen at random from the population of 30 million cows, and suppose that 0 (or 1, or 2) are found to be infected.

$\longrightarrow$ How many of the 30 million total cows would we estimate to be infected?

$\longrightarrow$ What is the 95% confidence interval for the total number of infected cows?

No. infected		
Obs'd	Est'd	95% CI
0	0	0 – 2767
1	750	19 – 4178
2	1500	182 – 5418

The case X = n

Suppose $X \sim$ Binomial(n, p) and we observe X = n.

Upper limit of 95% confidence interval for p: $\rightarrow 1$

Lower limit of 95% confidence interval for p:

p_L such that

$$\begin{aligned} & \Pr(X \geq n \mid p = p_L) = 0.025 \\ \implies & \Pr(X = n \mid p = p_L) = 0.025 \\ \implies & (p_L)^n = 0.025 \\ \implies & p_L = \sqrt[n]{0.025} \end{aligned}$$

In the case n = 25 and X = 25, the 95% CI for p is (0.86, 1.00).

Large n and medium p

Suppose $X \sim$ Binomial(n, p).

$$E(X) = n\,p \qquad SD(X) = \sqrt{n\,p(1-p)}$$

$$\hat{p} = X/n \qquad E(\hat{p}) = p \qquad SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

For large n and medium p, $\longrightarrow$ $\hat{p} \sim \text{Normal}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

Use 95% confidence interval $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$\longrightarrow$ Unfortunately, this can behave poorly.

$\longrightarrow$ Fortunately, you can just calculate exact confidence intervals.