Introduction to Inference

Introduction to Statistics
Carl von Ossietzky Universität Oldenburg
Fakultät III - Sprach- und Kulturwissenschaften

Population and sample

- The entire group of individuals that we want information about is called the **population**.
- A **sample** is a part of the population that we actually examine in order to gather information.
- A **voluntary response sample** consists of people who choose themselves by responding to a general appeal. Voluntary response samples are biased because people with strong opinions, especially negative opinions, are most likely to respond.
- **Sampling** selects a part of a population of interest to represent the whole.

Population and sample

- The design of a study is **biased** if it systematically favors certain outcomes.
- A **simple random sample** (SRS) of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance to be the sample actually selected.
- A **parameter** is a number that describes the **population**. A parameter is a fixed number, but in practice we do not know its value.
- A **statistic** is a number that describes a **sample**. The value of a statistic is known when we have taken a sample, but it can change from sample to sample. We often use a statistic to estimate an unkown parameter.

Population and sample

- **Example**: 2500 adults in the US were asked whether they like to shop for clothes.
- 1100 respondents gave a positive answer:

$$\hat{p} = \frac{1100}{2500} = 0.44$$

- $\hat{p} = 0.44$ is a *statistics*, p is the corresponding parameter: the proportion of all US residents that would agree.
- We do not know parameter p, so we use statistics $\hat{p}$ to estimate it.
- **Sampling variability**: the value of the statistic $\hat{p}$ will vary from sample to sample.
- The **sampling distribution** of a statistic is the distribution of values taken by the statistics in alle possible samples of the same size from the same population.

Population and sample

- Distribution described by shape, center and spread:
  - **Shape** (histograms): is the distribution normal?
  - **Center** (mean): a statistic used to estimate a parameter which is **unbiased** if the mean of its sampling distribution is equal to the true value of the parameter being estimated. Use random sampling.
  - **Spread** (standard deviation): the larger the sample, the smaller the **variability of a statistic**.
- **Statistical inference**: use a fact about a sample to estimate the truth about the whole population. We call this **statistical inference** since we **infer** conclusions about the wider population from a sample of selected individuals.

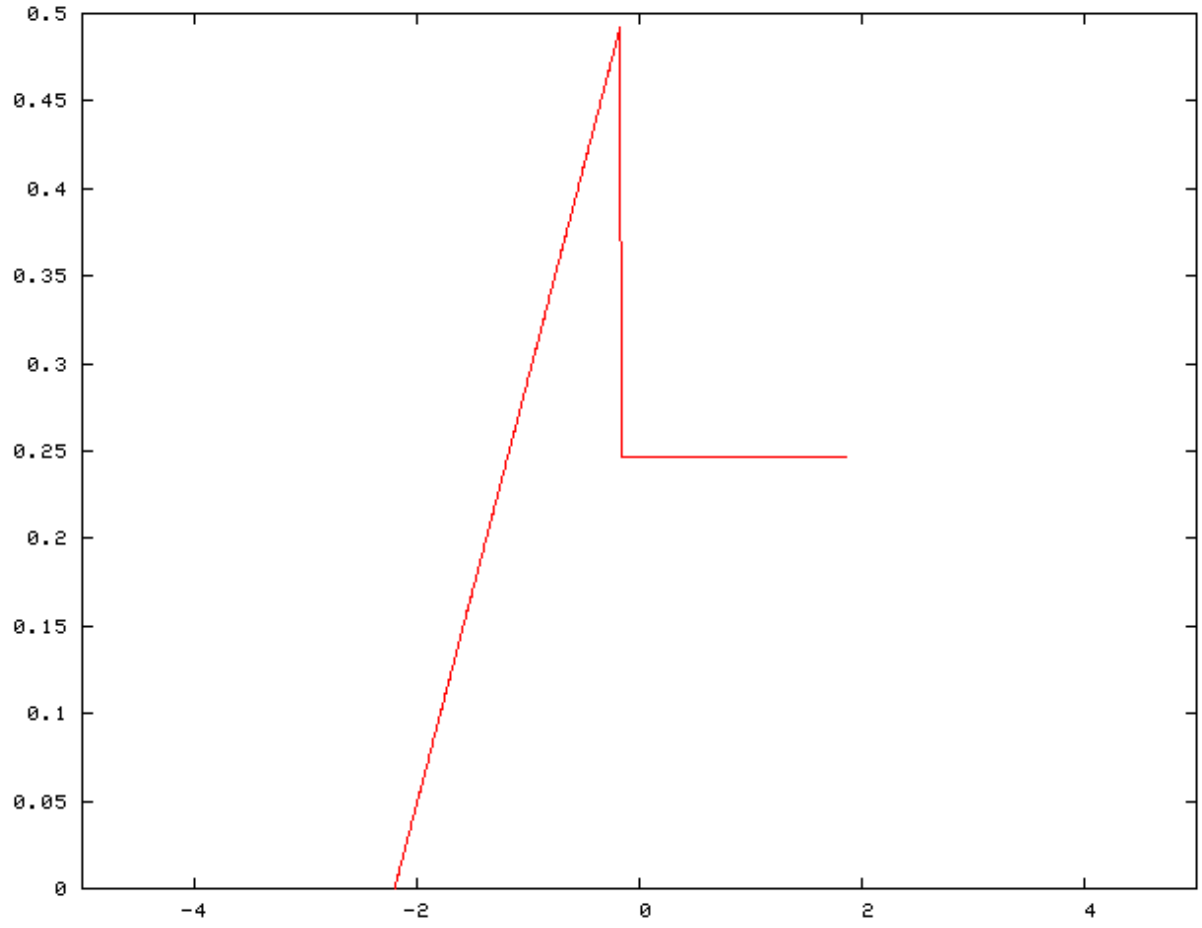
Sampling Distribution of a Sample Mean

- In example above: we looked at proportions: parameter p and statistics $\hat{p}$. They describe categorical data.
- For quantitative data looking at averages is very common: parameter μ and statistics $\overline{x}$.
- Averages are **less variable** than individual observations and averages are **more normal** distributed than individual observations.
- Let $\overline{x}$ be the mean of a simple random sample (SRS) of size n from a population having mean μ and standard deviation σ. The mean and standard deviation of $\overline{x}$ are:

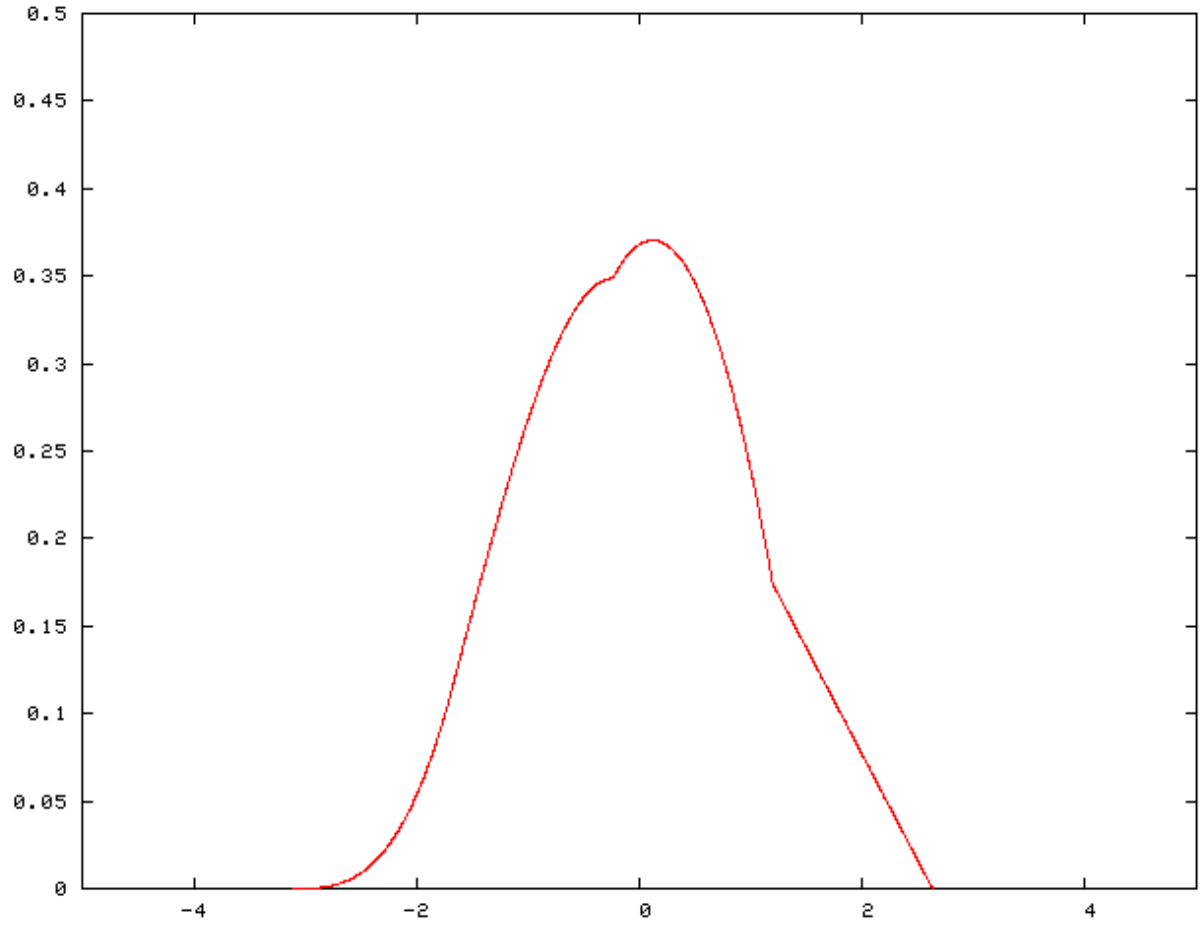
$$\mu_{\overline{x}} = \mu \qquad \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

Sampling Distribution of a Sample Mean

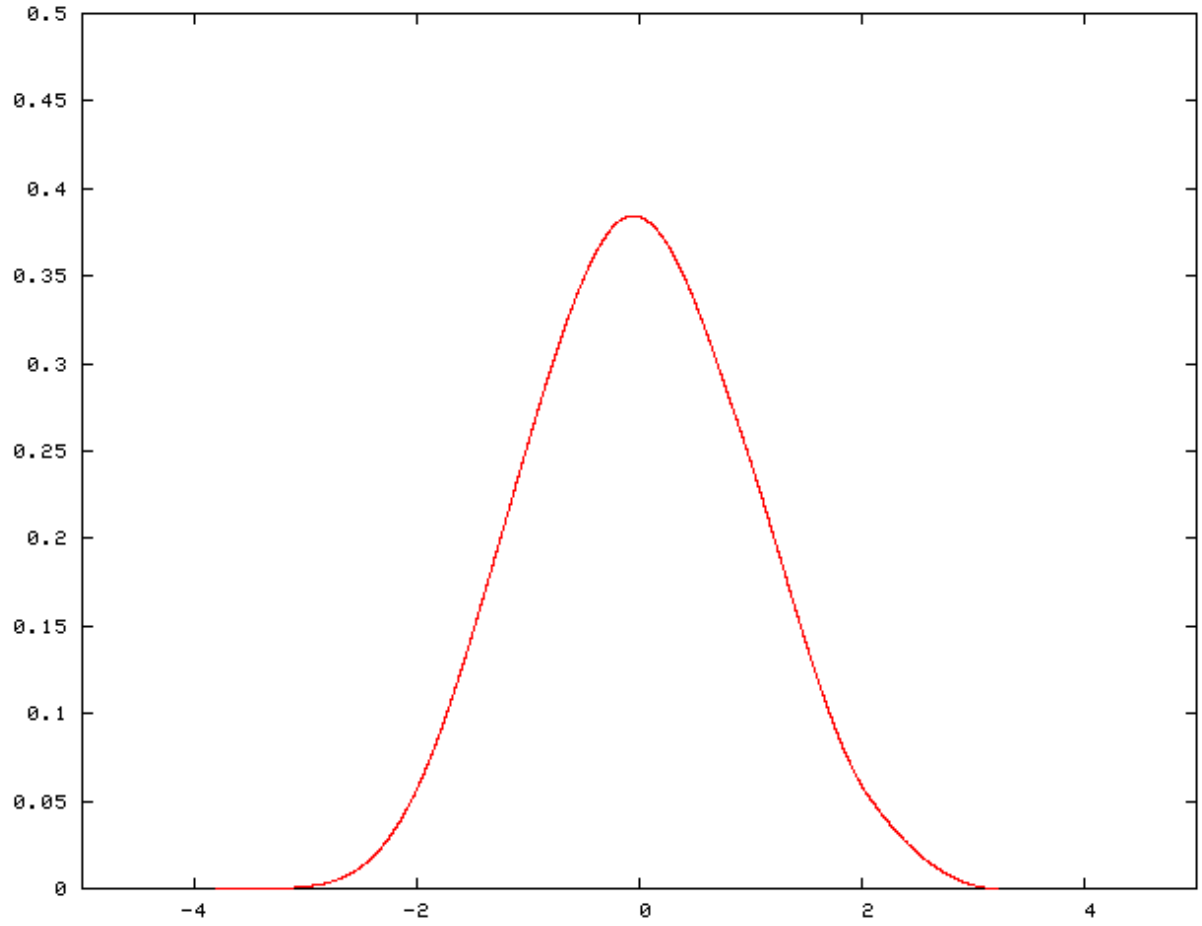
- **Central limit theorem**:
  - Suppose that a sample is obtained containing a large number of observations.
  - Each observation is randomly generated in a way that does not depend on the values of the other observations.
  - The average of the observed values in the sample is computed.
  - If this procedure is performed many times, the computed average will not always be the same each time.
  - The central limit theorem says that the computed values of the average will be distributed according to the normal distribution ("bell curve") when the sample size is large.
- If a population has the $N(\mu, \sigma)$ distribution, then the sample mean $\overline{x}$ of n independent observations has the $N(\mu, \sigma/\sqrt{n})$ distribution when n is large.



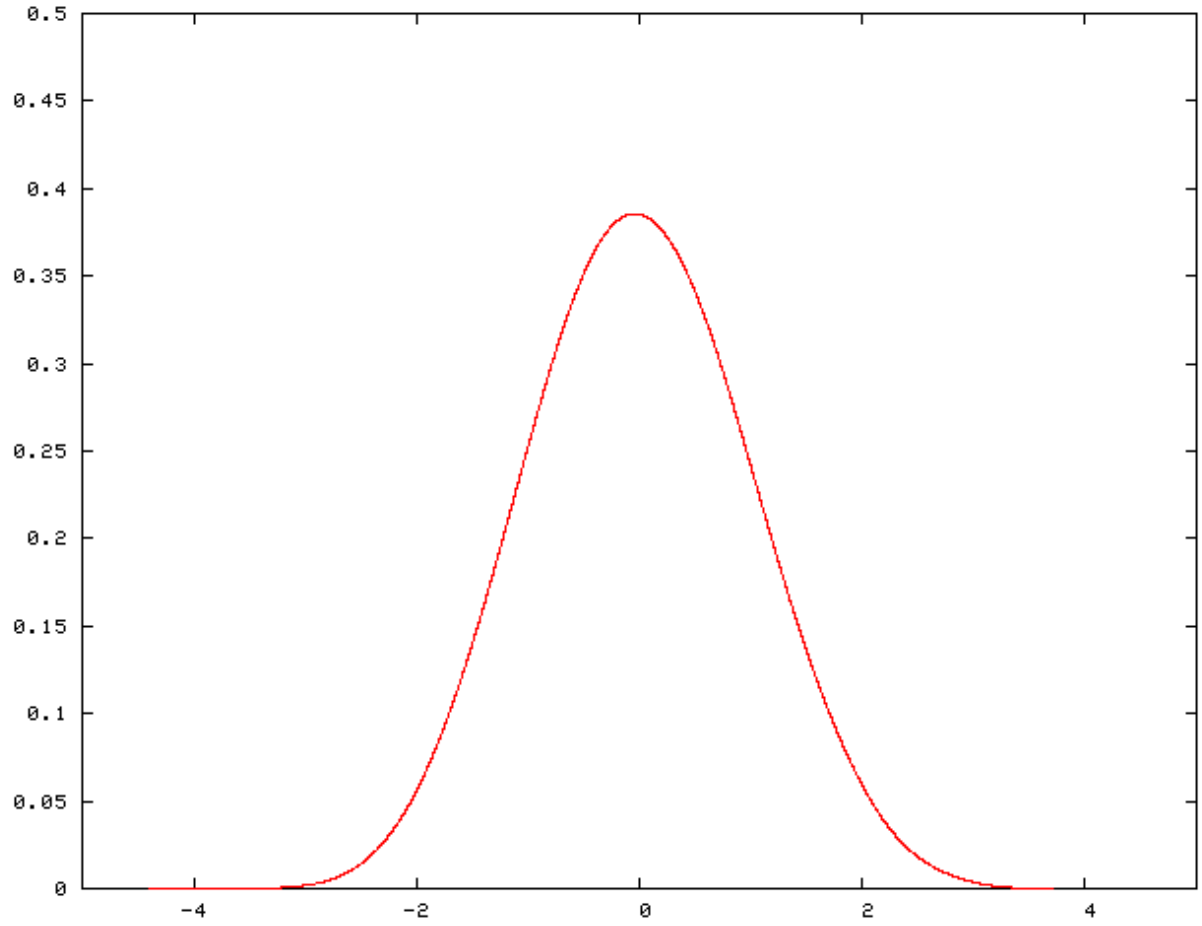
Density curve of individual observations. The mean of this distribution is 0 and its standard deviation is 1.4545.



The curve is the density curve of the sample mean of 2 observations.



The curve is the density curve of the sample mean of 3 observations.



The curve is the density curve of the sample mean of 4 observations.

Sampling Distribution of a Sample Mean

- Please have a look at:

  `http://www.statisticalengineering.com/central_limit_theorem.html`

- Have also a look at:

  `http://www.vias.org/simulations/simusoft_cenlimit.html`

  and download in the application (in English or German) and vary the sample size.

Sampling Distribution of a Sample Mean

- We return to the growth season example. We asked: what is the probability that a growth season will be shorter than 290 days given μ=204 days and σ=52.28 days. Answer: 0.9495 or 95%.
- New question: what is the chance that in a period of 2 years the growth season is **on average** shorter than 290 days?
- We express the problem in the standard scale of z-scores, where n=2:

$$z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{290 - 204}{\frac{52.28}{\sqrt{2}}} = 2.33$$

- According to the table with standard normal probabilities, e.g. at:

  `http://clas.sa.ucsb.edu/staff/binh/stdNormalTable.pdf`

  the area left of this value has a surface of 0.9901. So p=0.9901 or 99%.

Confidence intervals

- **Example**: reading ability of childeren is measured with a *Degree of Reading Power* (DRP) test.
- Scores for a sample of 44 thirdgrade students:

```
40 26 39 14 42 18 25 43 46 27 19 47 19 26 35 34 15 44 40 38 31 46
52 25 35 35 33 29 34 41 49 28 52 47 35 48 22 33 41 51 27 14 54 45
```

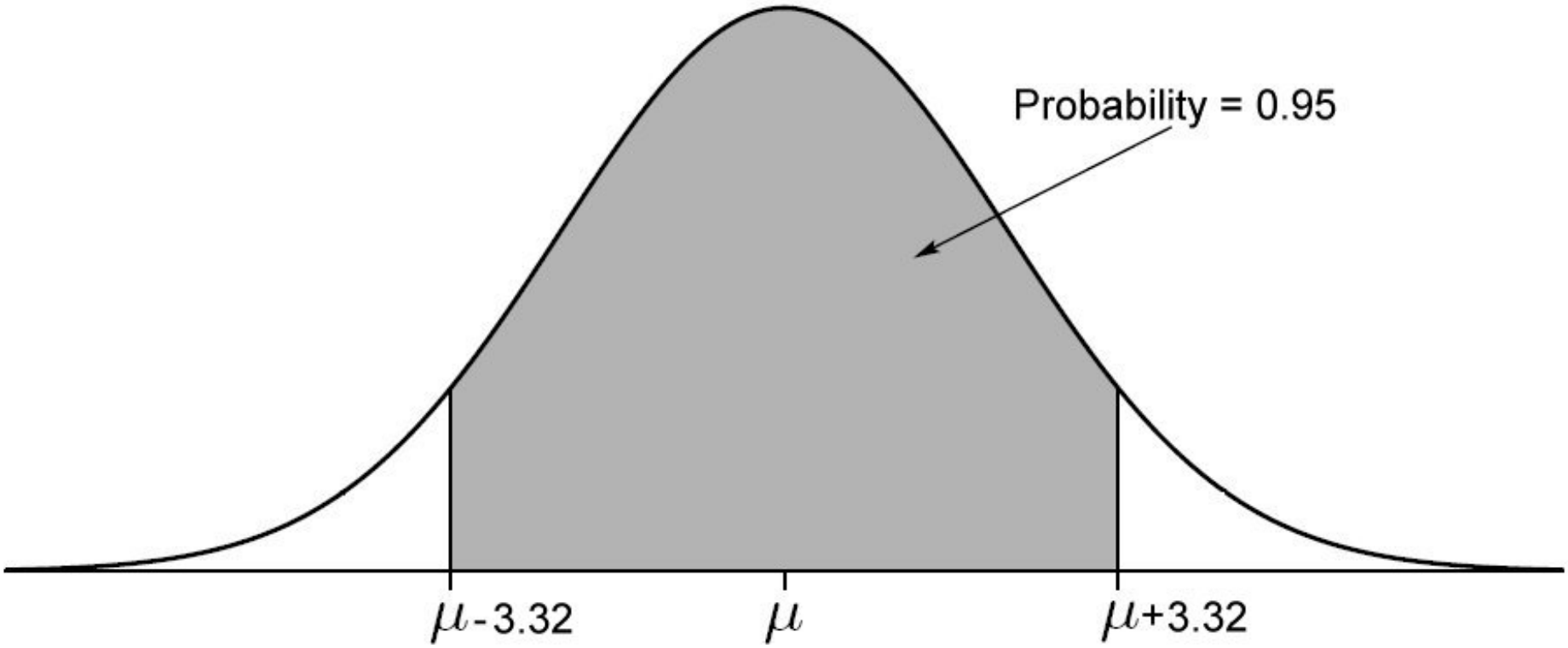
- Source: Maribeth Cassidy Schmitt, *The effects of an elaborated directed reading activity on the metacomprehension skills of thirdgraders*, PhD thesis, Purdue University 1987.
- $\overline{x}$=35.09 and s=11.19. Assume σ=11. We know the sample size n=44.

Confidence intervals

- At repeated sampling the sample mean $\overline{x}$ had a normal distribution centered at the unknown population mean μ and having standard deviation:

$$\sigma_{\overline{x}} = \frac{11}{\sqrt{44}} = 1.66$$

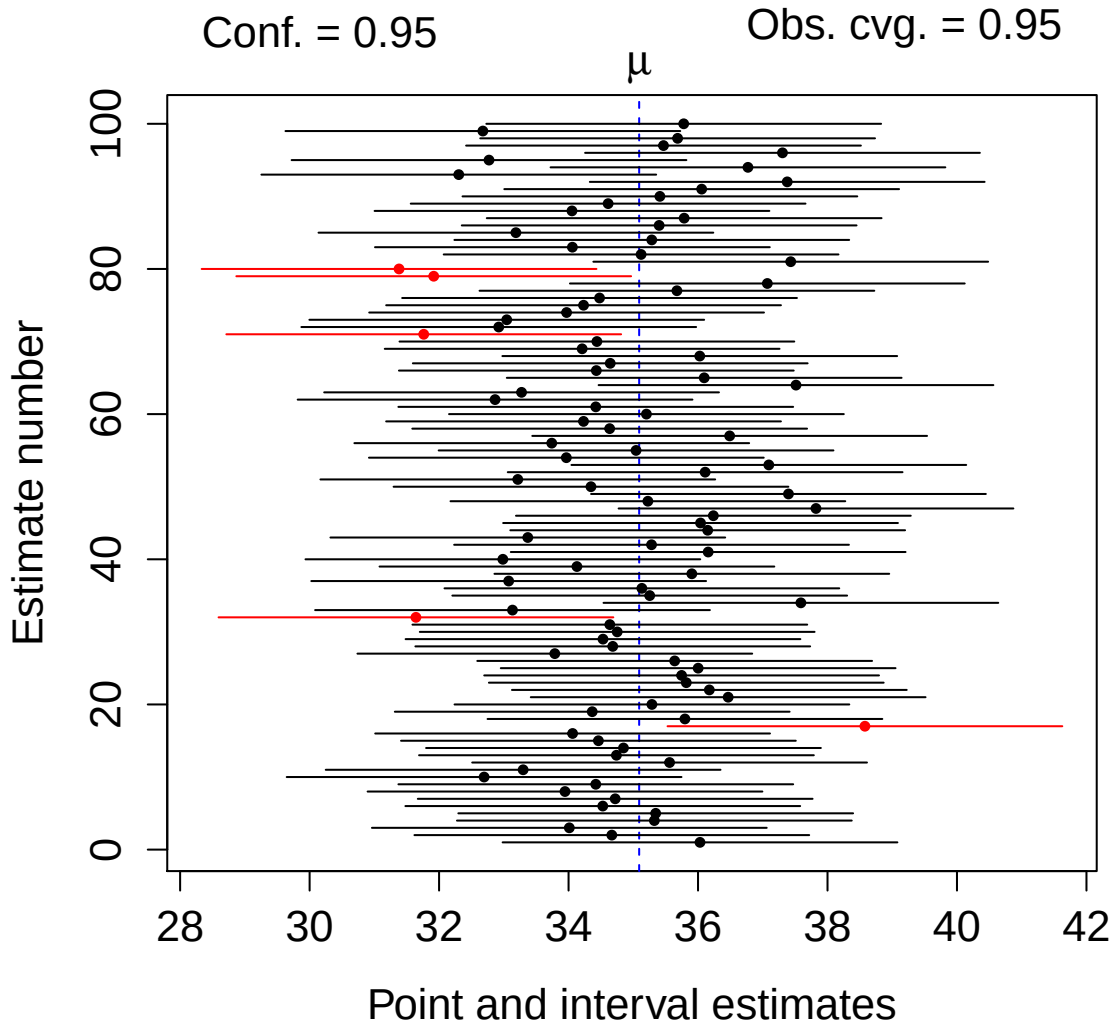
- The 68-95-99.7 rule says that the probability is about 0.95 that $\overline{x}$ will be within two standard deviations of the population mean score μ. Two standard deviations: $2 \times 1.66 = 3.32$.
- To say that $\overline{x}$ lies within 3.32 points of μ is the same as saying that μ is within 3.32 points of $\overline{x}$.
- So 95% of all samples will capture the true μ in the interval from $\overline{x} - 3.32$ to $\overline{x} + 3.32$.
- We are 95% *confident* that the unknown mean score for all children lies between 35.09-3.32 (31.75) and 35.09+3.32 (38.41).



Density curve of $\overline{x}$ with unkown μ. $\overline{x}$ lies within $\pm$ 3.32 of μ in 95% of all samples, so μ also lies within $\pm$ 3.32 of $\overline{x}$ in those samples. The interval of numbers between the values $\overline{x} \pm 3.32$ is called a 95% *confidence interval* for μ. The general form is: estimate $\pm$ margin of error.

Confidence intervals

- A level C **confidence interval** for a parameter is an interval computed from sample data by a method that has probability C of producing an interval containing the true value of the parameter.


Conf. = 0.95
Obs. cvg. = 0.95
μ
Estimate number
0
20
40
60
80
100
28
30
32
34
36
38
40
42
Point and interval estimates

Confidence intervals

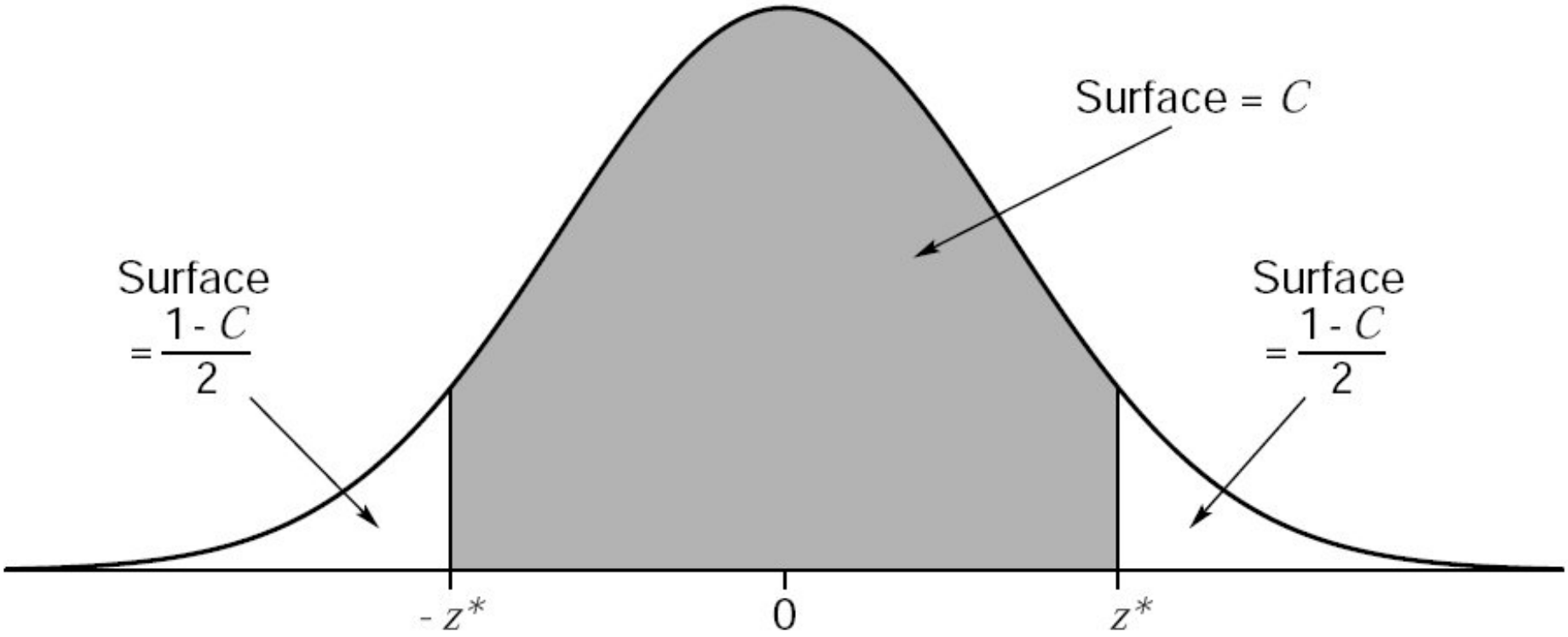
- To construct a level C confidence interval we first catch the central area C under a normal curve.
- We must find the number z^* such that any normal distribution has probability C within $\pm\ z^*$ standard deviations of its mean.
- There is probability C that $\overline{x}$ lies between:

$$\mu - z^*\frac{\sigma}{\sqrt{n}} \quad and \quad \mu + z^*\frac{\sigma}{\sqrt{n}}$$

- This is the same as saying that the unkown population mean μ lies between:

$$\overline{x} - z^*\frac{\sigma}{\sqrt{n}} \quad and \quad \overline{x} + z^*\frac{\sigma}{\sqrt{n}}$$

- The interval is exact when the population distribution is normal and is approximately correct for large n in other cases.



The area between $-z^*$ and z^* under the standard normal curve is C. If C=95%, than z^*=1.960 which means that the area C is found between -1.960 and 1.960. For finding $-z^*$ find in the table p=(1-C)/2=0.025. The corresponding z is -1.96, therefore $-z^*$=-1.960 and z^*=1.960.

Confidence intervals

- We return to the DRP example: $\overline{x}$=35.09. Assume σ=11. We know that sample size n=44.
- We calculate the C=95% confidence interval. For finding $-z^*$ find in the table p=(1-C)/2=0.025. The corresponding z is -1.96. Therefore $-z^*$=-1.960 and z^*=1.960. Then the unknown population mean μ lies between:

$$\overline{x} - z^* \frac{\sigma}{\sqrt{n}} \quad and \quad \overline{x} + z^* \frac{\sigma}{\sqrt{n}}$$

is:

$$35.09 - 1.960 \frac{11}{\sqrt{44}} \quad and \quad 35.09 + 1.960 \frac{11}{\sqrt{44}}$$

is:

$$31.84 \quad and \quad 38.34$$

- We are 95% *confident* that the unknown mean score for all children lies between 31.84 and 38.34.

Confidence intervals

- It would be nice to get a narrow interval with high confidence.
- Other things being equal, the width of a confidence interval decreases as
  - the confidence level C decreases,
  - the sample size n increases, and
  - the population standard deviation decreases.
- Question: we would like to find a confidence interval with a particular width and confidence. We want to find the interval on the basis of a sample. How large should the sample be?

Confidence intervals

- The sample size required to obtain a confidence interval of specified margin of error m for a normal mean is:

$$n = \left(\frac{z^* \sigma}{m}\right)^2$$

- The **margin of error** m expresses the error caused by observing a sample instead of the whole population. It defines the width of a confidence interval: the lower boundary is $\overline{x} - m$ and the upper boundary is $\overline{x} + m$.
- Example: we want to find a 99% confidence interval with $m = 2$ for the DRP scores. From the table we can derive z^* for C=0.99. We find z^*=2.576. Required sample size:

$$n = \left(\frac{z^* \sigma}{m}\right)^2 = \left(\frac{2.576 \times 11}{2}\right)^2 = 200.73$$

- Always round up to the next higher whole number when finding n because this will give us a smaller margin of error. So the sample should contain 201 scores.

Confidence intervals

Some cautions:

- The data should be an SRS.
- Because $\overline{x}$ is not resistant, search for outliers and try to correct them or justify their removal.
- If the sample size is small, examine your data carefully for skewness and other signs of nonnormality. When $n \geq 15$, the confidence level is not greatly disturbed by nonnormal populations unless extreme outliers or quite strong skewness are present.
- The standard deviation σ must be known, but this is an unrealistic requirement.
- A 95% confidence interval means that we are 95% confident that μ lies between the lower and upper boundary. The boundaries are calculated by a method that gives correct results in 95% of all possible samples. It does **not** say that the probability is 95% that the true mean falls between the lower and upper boundary.

Significance tests

- A **significance test** is a formal procedure for comparing observed data with a hypothesis whose truth we want to assess.
- Example: are Frisian dialects more distant to standard Dutch than dialects in the Dutch dialect area are on average?
- For 34 Frisian dialects we compare the pronunciation of 125 words with the pronunciation of the corresponding Dutch words.
- The more differences in pronunciation, the larger the distance between a dialect and standard Dutch. Distances are measured in percentages.
- Assume that the population of all Dutch dialect distances – measured in the same way – has $\mu = 30$ and $\sigma = 20$.
- The Frisian sample has a distance of 38.7%. Different Frisian samples would give different averages. Is the difference meaningful? Use significance test to find out!

The sample of 34 Frisian dialects have an average pronunciation distance of 38.7% to standard Dutch.

Significance tests

- Is the difference between the population mean (30%) and the sample mean (38.7%) reasonable if, in fact, the Frisian sample is taken from a population with a mean of 30%?
- **First step** in a test of significance: define null hypothesis.
- The statement being tested in a test of significance is called **null hypothesis**. The test of significance is designed to assess the strength of the evidence against the null hypothesis.
- Usually the null hypothesis is a statement of "no effect" or "no difference":

$$H_0\colon\ \mu = 30\%$$

Significance tests

- The **alternative hypothesis** is the statement we hope or suspect to be true instead of the null hypothesis.
- One-sided alternatives:

$$H_a\colon\ \mu > 30\%$$

and:

$$H_a\colon\ \mu < 30\%$$

- Two-sided alternative:

$$H_a\colon\ \mu \neq 30\%$$

Significance tests

- **Second step**: calculate the value of the test statistic on which the test will be based.
- A **test statistic** measures compatibility between the null hypothesis and the data of the sample. It estimates the parameter that appears in the hypothesis.
- When H_0 is true, we expect the estimate to take a value near the parameter value specified by H_0.
- Values of the estimate far from the parameter value specified by H_0 give evidence against H_0.
- Our test statistic measures the difference between the sample estimate and the hypothesized parameter in terms of standard deviations of the test statistic:

$$z = \frac{estimate - hypothesized\ value}{standard\ deviation\ of\ the\ estimate}$$

Significance tests

- If the null hypothesis is defined in terms of population mean(s), the estimate is the sample mean $\overline{x}$. Therefore the test statistic is:

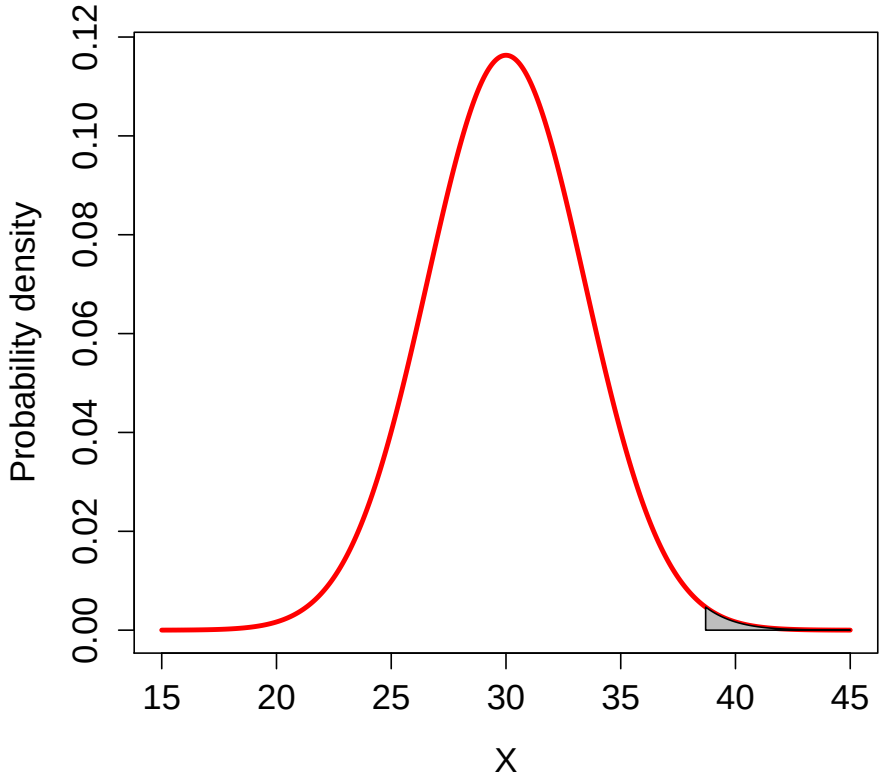
$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

- Recall: the central limit theorem says that $\overline{x}$ and z will be approximately normal when the sample size is large even if the population is not normal.
- We see that $15 \leq n < 40$. The sample does not have extreme outliers and is not quite strongly skewed, so we may assume a normal distribution.
- Assume σ=20, then:

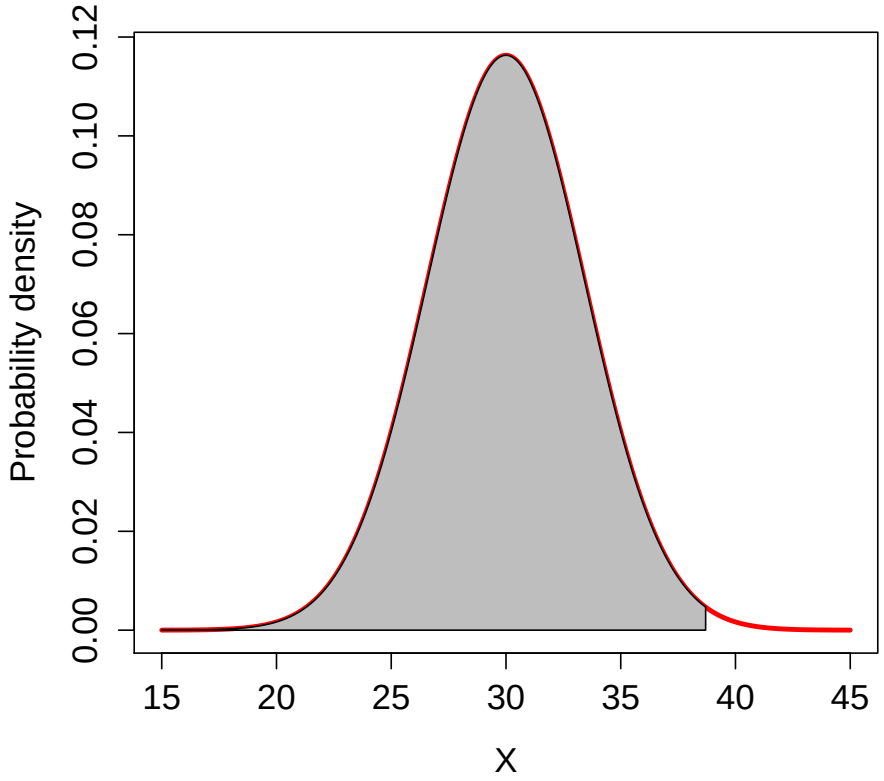
$$z = \frac{38.7 - 30}{20/\sqrt{34}} = 2.54.$$

Significance tests

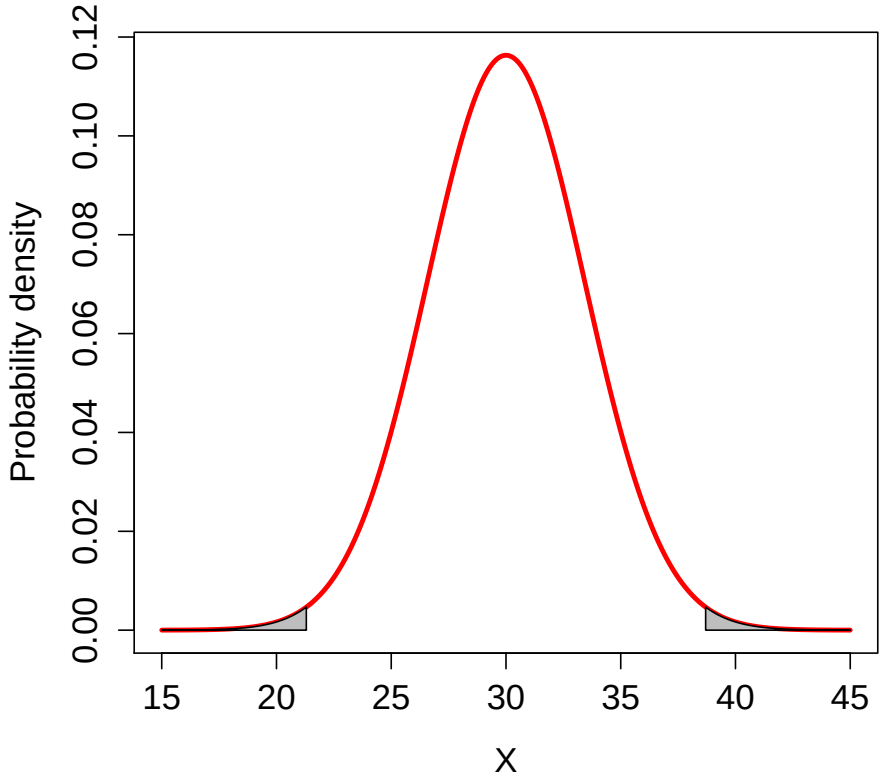
- **Third step**: find the P-value of the observed data.
- The probability, computed assuming that H_0 is true, that the test statistic would take a value as extreme or more extreme than the value actually observed is called P-**value** of the test. The smaller the P-value, the stronger the evidence against H_0 provided by the data.
- "Extreme" means "far from what we would expect if H_0 were true. The direction or directions that count as "far from what we would expect" are determined by H_a.
- If H_0 is true, then z is a single observation from the standard normal distribution $N(0, 1)$.
- The P-value is the probability of observing a value of z at least as extreme as the one we observed, z=2.54.



If H_a: $\mu > 30$, and z is positive, we consider the proportion right of z. According to the table, the probability for being extreme in positive direction is $P(Z \geq 2.54) = 1 - P(Z \leq 2.54) = 1 - 0.9945 = 0.0055$.



If H_a: $\mu < 30$, and z is negative – but this was not the case –, we consider the proportion left of z. According to the table, the probability for being extreme in the negative direction is $P(Z \leq 2.54) = 0.9945$.



If H_a: $\mu \neq 30$, we consider the proportion left of $-z$ and right of z. The probability for being extreme in one of the two directions is $P(Z \leq -2.54) + P(Z \geq 2.54) = 2P(Z \geq 2.54) = 2(0.0055) = 0.0110$.

Significance tests

- We can compare the P-values with a fixed value that we regard as decisive. The decisive value of P is called the **significance level**. It is denoted by α.
- If the P-value is as small or smaller than α, we say that the data are **statistically significant at level** α.
- If H_a: $\mu > 30$, we found $z = 2.54$ and $P(Z \geq 2.54) = 0.0055$, which is lower than α=0.05.
- We conclude that Frisian dialects are more distant to standard Dutch than dialects in the Dutch dialect area are on average.
- Significances at the α=0.05 level are indicated by *, at 0.01 by '**', at 0.001 by '****'.

Two-sided significance tests and confidence intervals

- A two-sided test at significance level α can be carried out directly from a confidence interval with confidence $C = 1 - \alpha$.
- If we choose $\alpha = 0.05$, than $C = 1 - 0.05 = 0.95$ or 95%. From the table we can find z^* for C=0.95. We find z^*=1.960.
- The lower and upper boundary of the interval are:

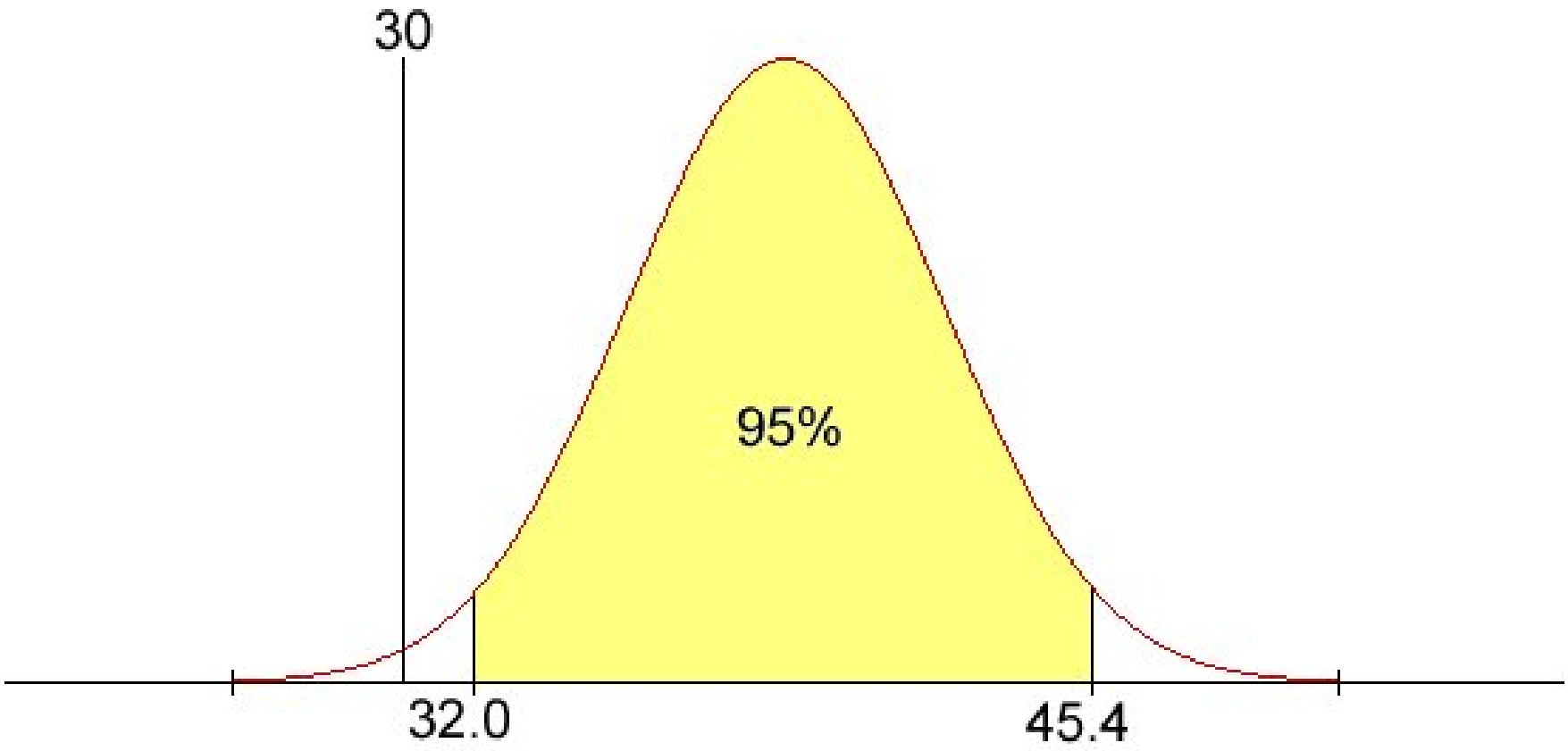
$$\overline{x} - z^* \frac{\sigma}{\sqrt{n}} \quad and \quad \overline{x} + z^* \frac{\sigma}{\sqrt{n}}$$

is:

$$38.7 - 1.960 \frac{20}{\sqrt{34}} \quad and \quad 38.7 + 1.960 \frac{20}{\sqrt{34}}$$

is:

$$32.0 \quad and \quad 45.4$$



Values of μ falling outside a 95% confidence interval can be rejected at the 5% significance level; values falling inside the interval cannot be rejected. $\mu = 30$ falls outside, so we reject H_0 and accept H_a.

Use and abuse of tests

- P-values are more informative than the reject-or-not result of a fixed level α test.
- Beware of placing too much weight on traditional values of α, such as α=0.05.
- Very small effects can be highly significant (small P), especially when a test is based on a large sample.
- A stastistic effect need not be practically important. Make graphs and calculate confidence intervals.
- Consider the **effect size**: an objective and (usually) standardized measure of the magnitude of observed effect (Field (2012)).
- Lack of significance does not imply that H_0 is true, especially when the test has low power.
- Significance tests are not always valid. Faulty data collection and outliers in the data can invalidate a test.

Power of tests

- The probability that a fixed level α significance test will reject H_0 when a particular alternative value of the parameter is true is called the **power** of the test to detect the alternative.
- Increase power by: increasing α, gives more extreme hypothesis, increase sample size, decreases σ.
- If we reject H_0 (accept H_a) when in fact H_0 is true, this is a **Type I error**. If we accept H_0 (reject H_a) when in fact H_a is true, this is a **Type II error**.

H_0	true	false
accepted	correct	type II error
rejected	type I error	correct