Estimation

The ASTA team

Contents

1 Estimation

1.1 Aim of statistics

- Statistics is all about “saying something” about a population.
- Typically, this is done by taking a random sample from the population.
- The sample is then analysed and a statement about the population can be made.
- The process of making conclusions about a population from analysing a sample is called **statistical inference**.

1.2 Random sampling schemes

Simple sampling(explained in Agresti section 2.2): Each experimental unit(fex persons) has the same probability of being selected(fex for an interview)

Other strategies for obtaining a random sample from the target population are explained in Agresti section 2.4:

- Systematic sampling
- Stratified sampling
- Cluster sampling
- Multistage sampling
- ...

1.3 Point and interval estimates

- We want to study hypotheses for population parameters, e.g. the mean μ and the standard deviation σ.
  - If μ is e.g. the mean waiting time in a queue, then it might be relevant to investigate whether it exceeds 2 minutes.
- Based on a sample we make a **point estimate** which is a guess of the parameter value.
  - For instance we have used $\bar{y}$ as an estimate of μ and s as an estimate of σ.
- We often want to supplement the point estimate with an **interval estimate** (also called a **confidence interval**). This is an interval around the point estimate, in which we are confident (to a certain degree) that the population parameter is located.
- The parameter estimate can then be used to investigate our hypothesis.

1.4 Point estimators: Bias

- If we want to estimate the population mean μ we have several possibilities e.g.
  - the sample mean $\bar{y}$
  - the average y_T of the sample upper and lower quartiles
- Advantage of y_T: Very large/small observations have little effect, i.e. it has practically no effect if there are a few errors in the data set.
- Disadvantage of y_T: If the distribution of the population is skewed, i.e. asymmetrical, then y_T is **biased**, meaning that in the long run this estimator systematically over or under estimates the value of μ.
- Generally we prefer that an estimator is **unbiased**, i.e. its distribution is centered around the true parameter value.
- Recall that for a sample from a population with mean μ, the sample mean $\bar{y}$ also has mean μ. That is, $\bar{y}$ is an unbiased estimate of the population mean μ.

1.5 Point estimators: Consistency

- From previous lectures we know that the standard error of $\bar{y}$ is $\frac{\sigma}{\sqrt{n}}$,
  - i.e. the standard error decrease when the sample size increase.
- In general an estimator with this property is callled **consistent**.
- y_T is also a consistent estimator, but has a variance that is greater than $\bar{y}$.

1.6 Point estimators: Efficiency

- Since the variance of y_T is greater than the variance of $\bar{y}$, $\bar{y}$ is preferred.
- In general we prefer the estimator with the smallest possible variance.
  - This estimator is said to be **efficient**.
- $\bar{y}$ is an efficient estimator.

1.7 Notation

- The symbol ˆ above a parameter is often used to denote a (point) estimate of the parameter. We have looked at an
  - estimate of the population mean: $\hat{\mu} = \bar{y}$
  - estimate of the population standard deviation: $\hat{\sigma} = s$
- When we observe a 0/1 variable, which e.g. is used to denote yes/no or male/female, then we will use the notation
$$\pi = P(Y = 1)$$
  for the proportion of the population with the property $Y = 1$.
- The estimate $\hat{\pi} = (y_1 + y_2 + \ldots + y_n)/n$ is the relative frequency of the property $Y = 1$ in the sample.

1.8 Confidence Interval

- The general definition of a confidence interval for a population parameter is as follows:
  - A **confidence interval** for a parameter is constructed as an interval, where we expect the parameter to be.
  - The probability that this construction yields an interval which includes the parameter is called the **confidence level** and it is typically chosen to be 95%.
  - (1-confidence level) is called the **error probability** (in this case $1 - 0.95 = 0.05$, i.e. 5%).
- In practice the interval is often constructed as a symmetric interval around a point estimate:
  - **point estimate±margin of error**
  - Rule of thumb: With a margin of error of 2 times the standard error you get a confidence interval, where the confidence level is approximately 95%.
  - I.e: **point estimate ± 2 x standard error** has confidence level of approximately 95%.

1.9 Confidence interval for proportion

- Consider a population with a distribution where the probability of having a given characteristic is π and the probability of not having it is $1 - \pi$.
- When *no/yes* to the characteristic is denoted 0/1, i.e. y is 0 or 1, the distribution of y have a standard deviation of:
$$\sigma = \sqrt{\pi(1 - \pi)}.$$
  That is, the standard deviation is not a “free” parameter for a 0/1 variable as it is directly linked to the probability π.
- With a sample size of n the standard error of $\hat{\pi}$ will be (since $\hat{\pi} = \frac{\sum_{i=1}^{n} y_i}{n}$):
$$\sigma_{\hat{\pi}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\pi(1 - \pi)}{n}}.$$

- We do not know π but we insert the estimate and get the **estimated standard error** of $\hat{\pi}$:

$$se = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}.$$

- The rule of thumb gives that the interval

$$\hat{\pi} \pm 2\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

has confidence level of approximately 95%. I.e., before the data is known the random interval given by the formula above has approximately 95% probability of covering the true value π.

1.9.1 Example: Point and interval estimate for proportion

- Now we will have a look at a data set concerning votes in Chile. Information about the data can be found here.

```
Chile <- read.delim("https://asta.math.aau.dk/datasets?file=Chile.txt")
```

- We focus on the variable `sex`, i.e. the gender distribution in the sample.

```
library(mosaic)
tally( ~ sex, data = Chile)
```

```
## sex
##    F    M
## 1379 1321
```

```
tally( ~ sex, data = Chile, format = "prop")
```

```
## sex
##         F         M
## 0.5107407 0.4892593
```

- Unknown population proportion of females (F), π.
- Estimate of π: $\hat{\pi} = \frac{1379}{1379+1321} = 0.5107$
- Rule of thumb : $\hat{\pi} \pm 2 \times se = 0.5107 \pm 2\sqrt{\frac{0.5107(1-0.5107)}{1379+1321}} = (0.49, 0.53)$ is an approximate 95% confidence interval for π.

1.9.2 Example: Confidence intervals for proportion in R

- **R** automatically calculates the confidence interval for the proportion of females when we do a so-called hypothesis test (we will get back to that later):

```r
prop.test( ~ sex, data = Chile, correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  Chile$sex  [with success = F]
## X-squared = 1.2459, df = 1, p-value = 0.2643
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4918835 0.5295675
## sample estimates:
##         p
## 0.5107407
```

- The argument `correct = FALSE` is needed to make **R** use the "normal" formulas as on the slides and in the book. When `correct = TRUE` (the default) a mathematical correction which you have not learned about is applied and slightly different results are obtained.

1.10 General confidence intervals for proportion

- Based on the central limit theorem we have:

$$\hat{\pi} \approx N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

  if $n\hat{\pi}$ and $n(1-\hat{\pi})$ both are at least 15.
- To construct a confidence interval with (approximate) confidence level $1-\alpha$:
  1) Find the socalled **critical value** z_{crit} for which the upper tail probability in the standard normal distribution is $\alpha/2$.
  2) Calculate $se = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$
  3) Then $\hat{\pi} \pm z_{crit} \times se$ is a confidence interval with confidence level $1-\alpha$.

1.10.1 Example: Chile data

Compute for the `Chile` data set the 99% and 95%-confidence intervals for the probability that a person is female:

- For a 99%-confidence level we have $\alpha = 1\%$ and
  1) z_{crit}=`qdist("norm", 1 - 0.01/2)`=2.576.
  2) We know that $\hat{\pi} = 0.5107$ and $n = 2700$, so $se = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = 0.0096$.
  3) Thereby, a 99%-confidence interval is: $\hat{\pi} \pm z_{crit} \times se = (0.4859, 0.5355)$.
- For a 95%-confidence level we have $\alpha = 5\%$ and
  1) z_{crit}=`qdist("norm", 1 - 0.05/2)`=1.96.
  2) Again, $\hat{\pi} = 0.5107$ and $n = 2700$ and so $se = 0.0096$.
  3) Thereby, we find as 95%-confidence interval $\hat{\pi} \pm z_{crit} \times se = (0.4918, 0.5295)$ (as the result of `prop.test`).

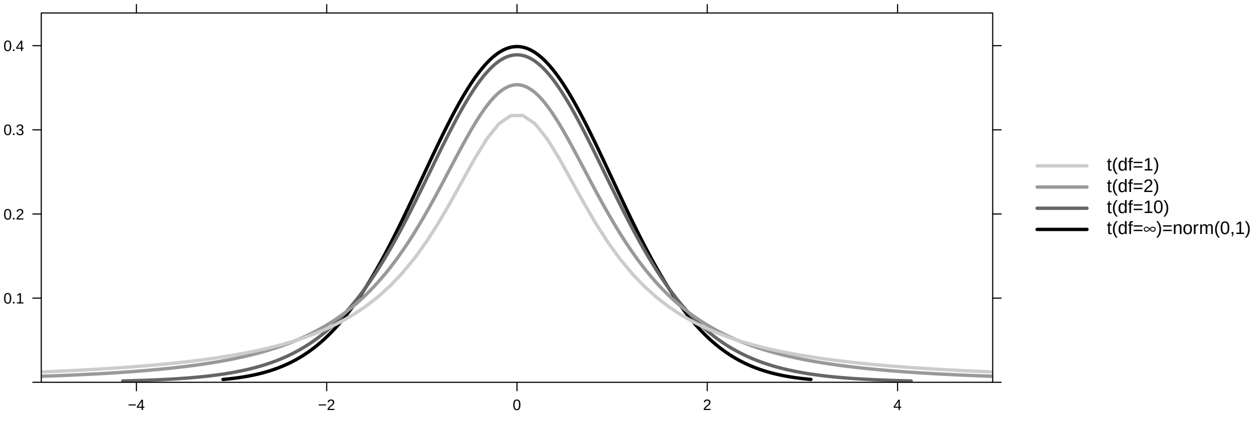
1.11 Confidence Interval for mean - normally distributed sample

- When it is reasonable to assume that the population distribution is normal we have the **exact** result
$$\bar{y} \sim \texttt{norm}(\mu, \frac{\sigma}{\sqrt{n}}),$$
i.e. $\bar{y} \pm z_{crit} \times \frac{\sigma}{\sqrt{n}}$ is not only an approximate but rather an exact confidence interval for the population mean, μ.
- In practice we **do not know** σ and instead we are forced to apply the sample standard deviation s to find the **estimated standard error** $se = \frac{s}{\sqrt{n}}$.
- This extra uncertainty, however, implies that an exact confidence interval for the population mean μ cannot be constructed using the z-score.
- Luckily, an exact interval can still be constructed by using the so-called ***t*-score**, which apart from the confidence level depends on the **degrees of freedom**, which are $df = n - 1$. That is the confidence interval now takes the form
$$\bar{y} \pm t_{crit} \times se.$$

1.12 t-distribution and t-score

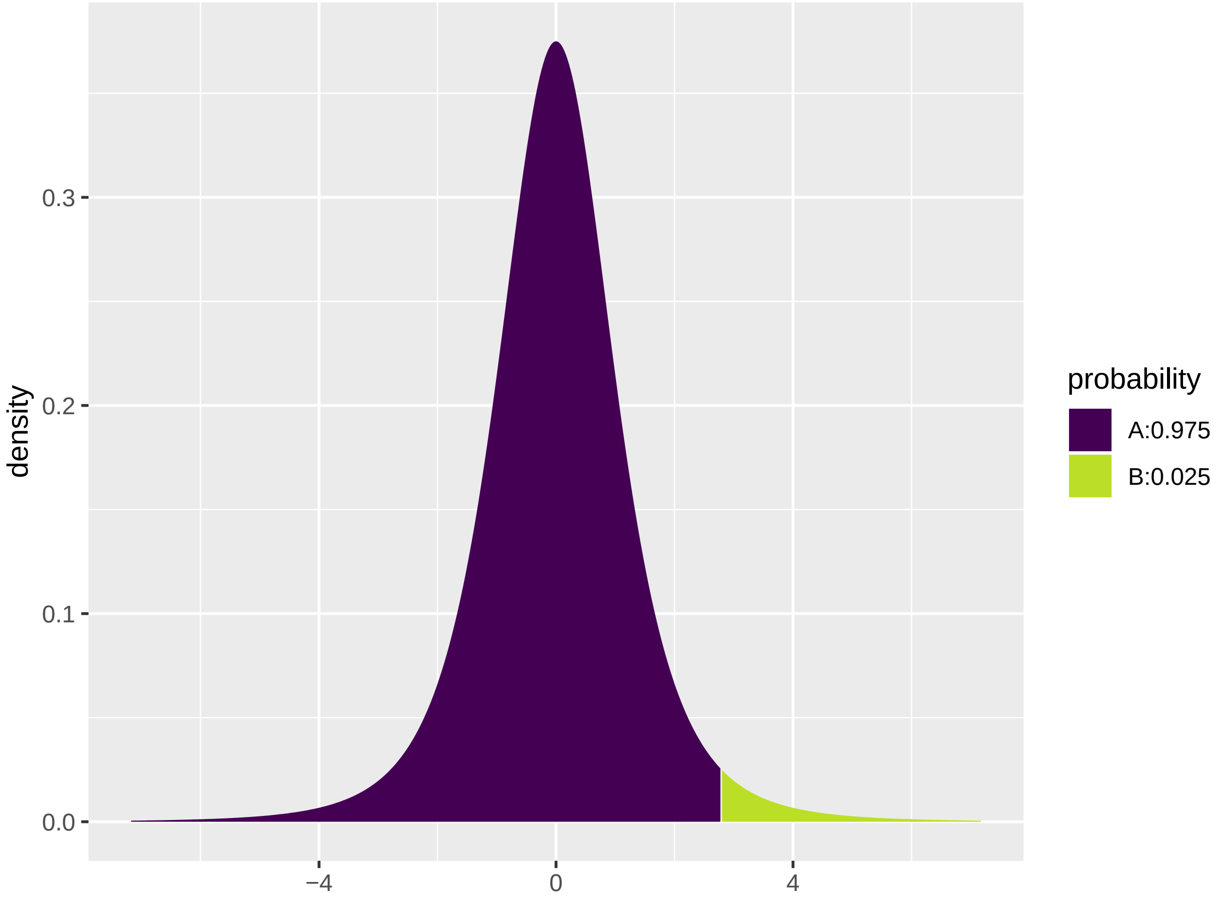
- Calculation of t-score is based on the ***t*-distribution**, which is very similar to the standard normal z-distribution:
  - it is symmetric around zero and bell shaped, but
  - it has "heavier" tails and thereby
  - a slightly larger standard deviation than the standard normal distribution.
  - Further, the t-distribution's standard deviation decays as a function of its **degrees of freedom**, which we denote df.
  - and when df grows the t-distribution approaches the standard normal distribution.

The expression of the density function is of slightly complicated form and will not be stated here, instead the t-distribution is plotted below for $df = 1, 2, 10$ and ∞.



1.13 Calculation of t-score in R

```
qdist("t", p = 1 - 0.025, df = 4)
```



```
## [1] 2.776445
```

- We seek the quantile (i.e. value on the x-axis) such that we have a given **right tail** probability. This is the critical t-score associated with our desired level of confidence.
- To get e.g. the t-score corresponding to a right tail probability of 2.5 % we have to look up the 97.5 % quantile using `qdist` with `p = 1 - 0.025` since `qdist` looks at the area to the **left hand side**.
- The degrees of freedom are determined by the sample size; in this example we just used df = 4 for illustration.
- As the t-score giving a right probability of 2.5 % is 2.776 and the t-distribution is symmetric around 0, we have that an observation falls between -2.776 and 2.776 with probability 1 - 2 · 0.025 = 95 % for a t-distribution with 4 degrees of freedom.

1.14 Example: Confidence interval for mean

- We return to the dataset `Ericksen` and want to construct a 95% confidence interval for the population mean μ of the variable `crime`.

```
Ericksen <- read.delim("https://asta.math.aau.dk/datasets?file=Ericksen.txt")
stats <- favstats( ~ crime, data = Ericksen)
stats
```

```
##  min Q1 median Q3 max     mean       sd  n missing
##   25 48     55 73 143 63.06061 24.89107 66       0
```

```
qdist("t", 1 - 0.025, df = 66 - 1, plot = FALSE)
```

```
## [1] 1.997138
```

- I.e. we have
  - $\bar{y} = 63.061$
  - $s = 24.891$
  - $n = 66$
  - $df = n - 1 = 65$
  - $t_{crit} = 1.997$.
- The confidence interval is $\bar{y} \pm t_{crit}\frac{s}{\sqrt{n}} = (56.942, 69.18)$
- All these calculations can be done automatically by **R**:

```
t.test( ~ crime, data = Ericksen, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  crime
## t = 20.582, df = 65, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  56.94162 69.17960
## sample estimates:
## mean of x
##  63.06061
```

1.15 Example: Plotting several confidence intervals in R

- We shall look at a built-in **R** dataset `chickwts`.
- `?chickwts` yields a page with the following information

> An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens. Newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement. Their weights in grams after six weeks are given along with feed types.

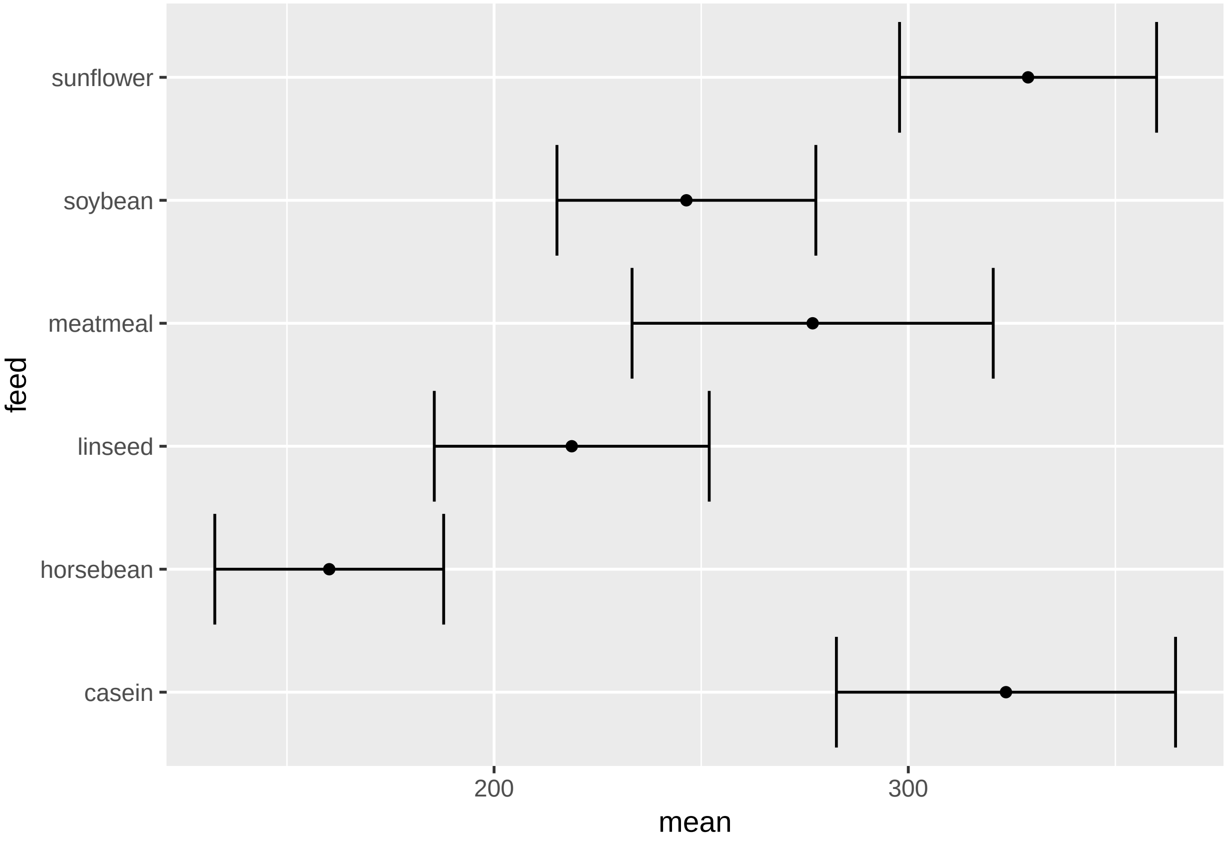
- `chickwts` is a data frame with 71 observations on 2 variables:
  - `weight`: a numeric variable giving the chick weight.
  - `feed`: a factor giving the feed type.
- Calculate a confidence interval for the mean weight for each feed separately; the confidence interval is from `lower` to `upper` given by `mean±tscore * se`:

```
cwei <- favstats( weight ~ feed, data = chickwts)
se <- cwei$sd / sqrt(cwei$n) # Standard errors
tscore <- qdist("t", p = .975, df = cwei$n - 1, plot = FALSE) # t-scores for 2.5% right tail probabilit
cwei$lower <- cwei$mean - tscore * se
cwei$upper <- cwei$mean + tscore * se
cwei[, c("feed", "mean", "lower", "upper")]
```

```
##        feed     mean    lower    upper
## 1    casein 323.5833 282.6440 364.5226
## 2 horsebean 160.2000 132.5687 187.8313
## 3   linseed 218.7500 185.5610 251.9390
## 4  meatmeal 276.9091 233.3083 320.5099
## 5   soybean 246.4286 215.1754 277.6818
## 6 sunflower 328.9167 297.8875 359.9458
```

- We can plot the confidence intervals as horizontal line segments using `gf_errorbarh`:

```
gf_errorbarh(feed ~ lower + upper, data = cwei) %>%
  gf_point(feed ~ mean)
```



2 Determining sample size

2.1 Sample size for proportion

- The confidence interval is of the form point estimate±estimated margin of error.

- When we estimate a proportion the margin of error is

$$M = z_{crit}\sqrt{\frac{\pi(1-\pi)}{n}},$$

where the critical z-score, z_{crit}, is determined by the specified confidence level.
- Imagine that we want to plan an experiment, where we **want to achieve a certain margin of error** M (and thus a specific width of the associated confidence interval).
- If we solve the equation above we see:
  - If we choose sample size $n = \pi(1-\pi)(\frac{z_{crit}}{M})^2$, then we obtain an estimate of π with margin of error M.
- If we do not have a good guess for the value of π we can use the worst case value $\pi = 50\%$. The corresponding sample size $n = (\frac{z_{crit}}{2M})^2$ ensures that we obtain an estimate with a margin of error, which is at the *most* M.

2.1.1 Example

- Let us choose $z_{crit} = 1.96$, i.e the confidence level is 95%.
- How many voters should we ask to get a margin of error, which equals 1%?
- Worst case is $\pi = 0.5$, yielding:

$$n = \pi(1-\pi)\left(\frac{z_{crit}}{M}\right)^2 = \frac{1}{4}\left(\frac{1.96}{0.01}\right)^2 = 9604.$$

- If we are interested in the proportion of voters that vote for “socialdemokratiet” a good guess is $\pi = 0.23$, yielding

$$n = \pi(1-\pi)\left(\frac{z_{crit}}{M}\right)^2 = 0.23(1-0.23)\left(\frac{1.96}{0.01}\right)^2 = 6804.$$

- If we instead are interested in “liberal alliance” a good guess is $\pi = 0.05$, yielding

$$n = \pi(1-\pi)\left(\frac{z_{crit}}{M}\right)^2 = 0.05(1-0.05)\left(\frac{1.96}{0.01}\right)^2 = 1825.$$

2.2 Sample size for mean

- The confidence interval is of the form point estimate±estimated margin of error.
- When we estimate a mean the margin of error is

$$M = z_{crit}\frac{\sigma}{\sqrt{n}},$$

where the critical z-score, z_{crit}, is determined by the specified confidence level.
- Imagine that we want to plan an experiment, where we **want to achieve a certain margin of error** M.
- If we solve the equation above we see:
  - If we choose sample size $n = (\frac{z_{crit}\sigma}{M})^2$, then we obtain an estimate with margin of error M.
- Problem: We usually do not know σ. Possible solutions:
  - Based on similar studies conducted previously, we make a qualified guess at σ.
  - Based on a pilot study a value of σ is estimated.