Statistical Inference

Bernhard Klingenberg

Institute of Statistics
Graz University of Technology
Steyrergasse 17/IV, 8010 Graz
www.statistics.tugraz.at

February 12, 2008

Outline

Estimation:

- Review of concepts
- Population vs. Sample
- Shape, Center, Spread
- Estimation of population parameters

Inference: Confidence Intervals and Hypothesis Tests for

- Proportions in one and two samples
- Means in one and two samples

Review

In statistics, any kind of information we want to obtain and any conclusions we want to draw are based on **data**
How do we get data?

- Often, we obtain data by taking a random sample from a large population and observe/measure various variables
- But see also the concept of Experiments

Different types of variables:

- Categorical and numerical variables
- Discrete and continuous variables

Review

Describing the distribution of variables:

- Depends on type of variable
- By means of graphics:
  - Bar Chart, Pie Chart, Histogram, Boxplot, Q-Q Plots, Scatterplot, ...
- By means of numerical summaries:
  - For categorical/discrete variables: Frequencies and relative frequencies (proportions)
  - For numerical/continuous variables: Mean, Median, Mode, Quartiles, Percentiles, Standard Deviation, IQR, Correlation, ...

Review

What do we look for?

- **Shape:** Symmetric or skewed distribution, number of modes (bimodal, etc.)
- **Center:** Where are most of the values located? Where does the distribution peak?
- **Spread:** How disperse (variable) are the values, what is the smallest, largest value?

Estimation

- The distribution (i.e., shape, center and spread) of a variable X in the population is unknown! (We call X a random variable)
- We wish to **estimate** some of its characteristics (e.g., the center) by taking a sample $X_1, X_2, \ldots, X_n$ of n observations or measurements of the variable
- We assume that all observations are independent and come from the same distribution (the true distribution of the variable in the population)
- This leads to independent and identically distributed (**iid**) observations

Estimation

- Notation: $X_i \overset{iid}{\sim} F$, $i = 1, \ldots, n$, where F is the true but unknown distribution function of the variable X.
- Example 1: Suppose F denotes the unknown distribution of the variable FVC in the population of all firemen.
- We say that the FVC values from the 79 firemen in our data set is a random (iid) sample from the true distribution.
- Based on this random sample, can we **estimate** the center (e.g., mean) of the FVC distribution, i.e., a "typical" FVC value for this population of firemen?
- Can we **estimate** the spread (e.g., standard deviation) of the distribution?

Estimation

- Example 2: Suppose F denotes the unknown distribution of the variable "region" in the population of all firemen.
- We say that the sample of region membership for the $n = 79$ firemen in our data set is an iid sample from the true distribution.
- Based on this random sample, can we **estimate** the true proportion of firemen from region 1 in the entire population of firemen?

Estimation

We take a sample from a population to learn about (i.e., estimate) population parameters such as:

Population Parameter:	Mean μ	Prop. π	Std. Dev. σ	Corr. ρ	Slope β
Estimate or Statistic:	$\hat{\mu} = \bar{x}$	$\hat{\pi} = p$	$\hat{\sigma} = s$	$\hat{\rho} = r$	$\hat{\beta} = b$

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean (average)
p is the sample proportion
$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ is the sample standard deviation
r is the sample correlation coefficient
b is the estimated slope in a linear regression model

Estimation

What are population parameters?

- Remember the random variable X and its distribution F.
- Population parameters specify F and help to describe X.

Most important one: **Expected value** $\mathrm{E}(X)$

- Describes a "typical" value; center/location
- Definition:

$$\mathrm{E}(X) = \begin{cases} \int_{-\infty}^{\infty} x f(x) dx & \text{continuous case} \\ \sum_{\text{all } x\text{'s}} x P(X = x) & \text{discrete case} \end{cases}$$

- We often refer to the expected value as μ

Estimation

Example 1 (cont.): Suppose the FVC distribution in the population of all firemen follows a normal model with mean $\mu = 5.8$ and standard deviation $\sigma = 0.8$.
Thus, $F \equiv N(\mu, \sigma^2) \equiv N(5.8, 0.8^2)$ and we write $X \sim N(5.8, 0.8^2)$, where X is the random variable denoting FVC.
What is the expected value of X?

$$\begin{aligned} \mathrm{E}(X) &= \int_{-\infty}^{\infty} x f(x) dx = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right\} dx \\ &= \cdots = \mu = 5.8 \end{aligned}$$

It's just what we called μ in the normal model!
Interpretation: We expect to see a FVC value of 5.8 when observing this random variable.

Estimation

Example 2 (cont.): Suppose the number of firemen from region 1 follows a Binomial model with true proportion of firemen from region 1 equal to $\pi = 41\%$. Among the $n = 79$ firemen sampled, how many do we expect to come from region 1?
Let $X = \#(\text{firemen from region 1})$, write $X \sim \text{Binomial}(79, 0.41)$
What is the expected value of X?

$$\mathrm{E}(X) = \sum_{x=0}^{n} xP(X = x) = \sum_{x=0}^{n} x \binom{n}{x} \pi^x (1-\pi)^{n-x} = \cdots = n\pi$$

$$\mathrm{E}(X) = \sum_{x=0}^{n} x \binom{79}{x} 0.41^x (1-0.41)^{79-x} = \cdots = 79 \times 0.41 = 32.39$$

Interpretation: From the 79 firemen, we expect 32.4 to be from region 1.

Estimation

Another important population parameter: **Variance** $\mathrm{var}(X)$

- Describes the variability (spread) of the random variable around its expected value
- Definition:

$$\mathrm{var}(X) = \begin{cases} \int_{-\infty}^{\infty} (x - \mathrm{E}(X))^2 f(x) dx & \text{continuous case} \\ \sum_{\text{all } x\text{'s}} (x - \mathrm{E}(X))^2 P(X = x) & \text{discrete case} \end{cases}$$

- We often refer to the variance as σ^2 and to the **standard deviation**, which is simply its square root, as σ.

Estimation

Example 1 (cont.): Remember $X \sim N(5.8, 0.8^2)$, where X was the random variable denoting FVC. What is the variance of X?

$$\begin{aligned} \mathrm{var}(X) &= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} dx \\ &= \cdots = \sigma^2 = 0.64 \end{aligned}$$

Its just what we called σ^2 in the normal model!
The more meaningful measure for spread is the standard deviation, which is simply $\sqrt{0.64} = 0.8$.
Interpretation: The spread of the FVC values around their mean is 0.8.
For bell-shaped distributions, about 68% of observations fall within one standard deviation of the mean, and 95% of observations fall within two standard deviation of the mean.

Estimation

Example 2 (cont.): Remember $X = \#(\text{firemen from region 1})$ and $X \sim \text{Binomial}(79, 0.41)$.
What is the spread in the number of firemen from region 1?

$$\begin{aligned}\text{var}(X) &= \sum_{x=0}^{n}(x-\mu)^2 P(X=x) = \sum_{x=0}^{n}(x-n\pi)^2 \binom{n}{x}\pi^x(1-\pi)^{n-x} \\ &= \cdots = n\pi(1-\pi) = 79 \times 0.41 \times (1-0.49) = 19.1\end{aligned}$$

Interpretation: The standard deviation of the number of firemen from region 1 is 4.4.

When sample size n is large and π around 0.5, then can use same rule of thumb as before (68%, 95%).

Estimation

Rules for expected values and variances:

$$\mathrm{E}(aX+b) = a\mathrm{E}(X)+b, \qquad \mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{E}(X_i)$$

$$\mathrm{var}(aX+b) = a^2\mathrm{var}(X), \qquad \mathrm{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{var}(X_i)$$

The last identity only holds if the X_i's are independent.

Estimation

Careful:

$$\mathrm{var}(X_1 - X_2) = \mathrm{var}(X_1) + \mathrm{var}(X_2) - 2\mathrm{cov}(\mathrm{X}_1, \mathrm{X}_2)$$

If X_1 and X_2 are independent

$$\begin{aligned} \mathrm{cov}(\mathrm{X}_1, \mathrm{X}_2) &= \mathrm{E}\Big((X_1 - \mathrm{E}(X_1)) \times (X_2 - \mathrm{E}(X_2))\Big) \\ &= \mathrm{E}(X_1 \times X_2) - \mathrm{E}(X_1) \times \mathrm{E}(X_2) = 0 \end{aligned}$$

Sum of two normal random variables is again normal:

$$X_1 \sim N(\mu_1, \sigma_1^2),\ X_2 \sim N(\mu_2, \sigma_2^2) \implies X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Estimation

- In the two examples, we assumed to know the true distribution of X exactly! This is unrealistic.
- Often, we know (or guess) the shape (e.g., normal, binomial, exponential), but not the values of the parameters that define the shape. We need to estimate them based on a random sample $X_1, \ldots, X_n$.
- What is a reasonable estimate of the expected value (i.e., the true mean μ or true proportion π?
  - **The sample average.**
- What is a reasonable estimate of the variance σ^2 and the standard deviation σ?
  - **The sample variance and the sample standard deviation.**

Estimation

We take a sample from a population to learn about (i.e., estimate) population parameters such as:

Population Parameter:	Mean μ	Prop. π	Std. Dev. σ	Corr. ρ	Slope β
Estimate or Statistic:	$\hat{\mu} = \bar{x}$	$\hat{\pi} = p$	$\hat{\sigma} = s$	$\hat{\rho} = r$	$\hat{\beta} = b$

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean (average)
p is the sample proportion
$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ is the sample standard deviation
r is the sample correlation coefficient
b is the estimated slope in a linear regression model

Estimation

- Why is the sample mean $\bar{x}$ a reasonable estimate of the expected value μ (the true population mean)?
- Let's see what is a typical value for the sample mean, i.e., let's find the expected value of the sample mean constructed from a random sample $X_1, \ldots, X_n$:

$$\mathrm{E}\left(\bar{X}\right) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

- Hence, a typical value for the sample mean is exactly the true population mean (and not something else)!
- We call the sample mean an **unbiased estimator** of the true population mean μ.
- Can we find the spread of the sample mean?

Estimation

$$\text{var}\left(\bar{X}\right) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}$$

- Hence, the standard deviation of the sample mean is

$$\sqrt{\text{var}(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

- The larger the sample size n, the smaller the spread of the sample mean $\bar{X}$.
- But not proportionally: taking 4 times as large a sample only reduces the spread (=precision) of $\bar{X}$ by a factor of 2!

Estimation

An important Fact: **The Central Limit Theorem**

- Let $X_1, \ldots, X_n$, be iid random variables (a random sample) with $\mathrm{E}(X_i) = \mu$ and $\mathrm{var}(X_i) = \sigma^2$, $i = 1, \ldots, n$. (No other assumptions about F are necessary, such as shape)
- Then, for n sufficiently large, the **sample mean** $\bar{X}$ follows a **normal distribution** with mean μ and variance σ^2/n. For short:

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

- Applied to 0/1 (Bernoulli) random variables, this means that the sample proportion P (which is just $\bar{X}$) follows a normal distribution with mean equal to the true proportion π and variance equal to $\pi(1-\pi)/n$. For short:

$$P \sim N(\pi, \pi(1-\pi)/n)$$

Estimation

- One last thing: We cannot compute the standard deviation of $\bar{X}$! It contains the population parameter σ.
- But, we can estimate σ^2 by s^2, and plug it into the formula.
- This gives the so called **standard error:** $s/\sqrt{n}$.
- That is, the standard error of the sample mean $\bar{X}$ is $s/\sqrt{n}$.
- Standard errors are the key components to measure precision of estimators such as the sample mean or the sample proportion.
- The magnitude of the standard error **reflects the uncertainty** we have in estimating a population parameter by using a random sample of size n.
- In fact, can we give an interval of plausible values of a population parameter based on a random sample?

Inference: Confidence Interval for π

- Example 2 (cont.) Based on the random sample of $n = 79$ firemen from the 2 regions, can we find a range of plausible values for the true proportion of firemen that are from region 1?
- Obviously, we can get a point estimate for the true proportion by finding the sample proportion of firemen from region 1 in our sample of 79 firemen.

```
> firemen <- read.table("firemen.dat", header=TRUE)
> attach(firemen)
> mean(region==1)
> [1] 0.4303797
```

  The sample proportion is $p = 0.43$, or 43%.
- But would we take another random sample of 79 new firemen, we would obtain a (slightly) different p.

Inference: Confidence Interval for π

- Simulation experiment: Assume we know that there are exactly 1000 firemen in the two regions, 410 of which are from region 1. We can use R to simulate a random sample of $n = 79$ firemen:

```
> region.all <- c(rep(1, 410), rep(2, 590))
> region.sample <- sample(region.all, 79, replace=TRUE)
> mean(region.sample==1)
> 0.3797468
```

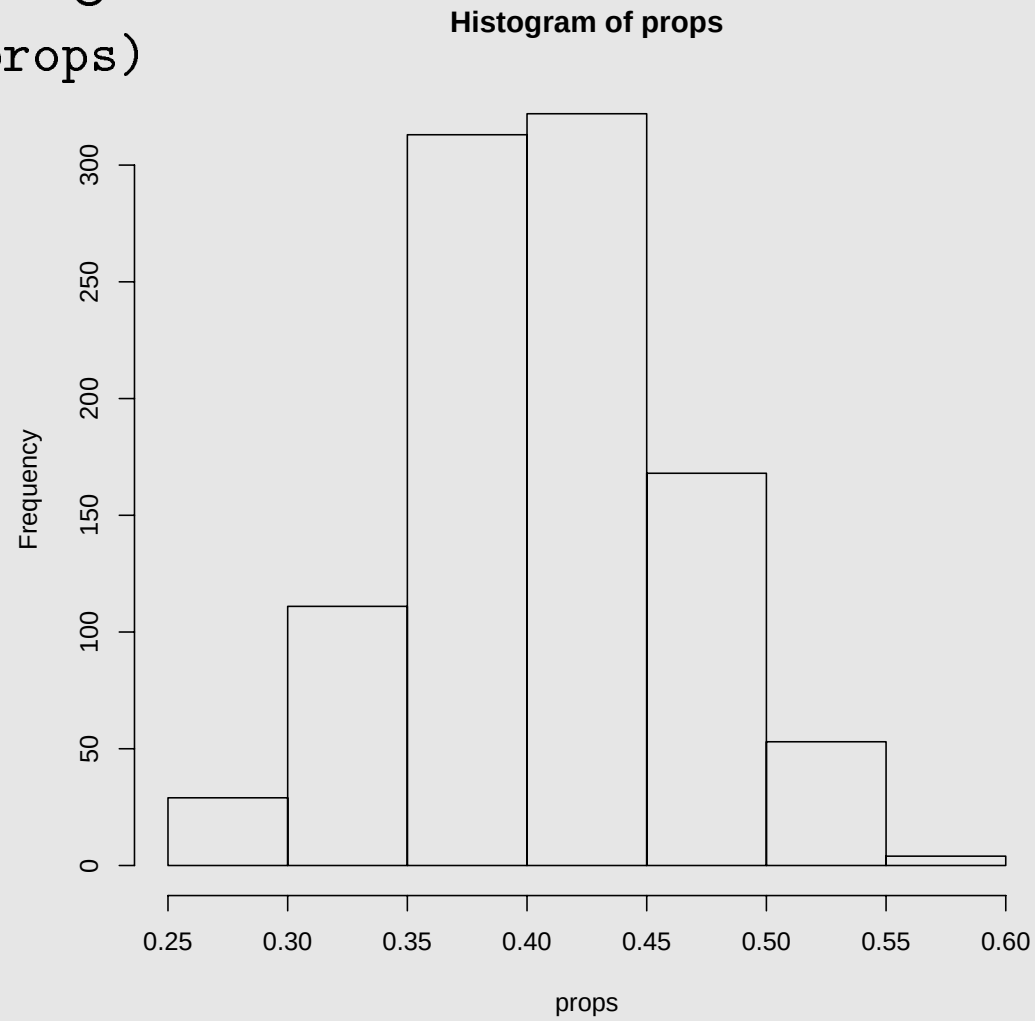
- Do this a number of times, and save each proportion

```
> props <- matrix(NA, 1000, 1)
> for (i in 1:1000) {
    region.sample <- sample(region.all, 79, replace=TRUE)
    props[i] <- mean(region.sample==1)
  }
```

Inference: Confidence Interval for π

- Plot a Histogram:

  ```
  > hist(props)
  ```



- This shows a range of plausible values

Inference: Confidence Interval for π

- Now, in reality we only have a **single sample**, and we don't know the **true proportion** of firemen from region 1. (That's precisely what we want to find out a range for!)
- But, can we estimate the spread of the histogram from just our single sample, without knowing the true proportion?
- Remember CLT (for large n): $P \sim N(\pi, \pi(1-\pi)/n)$.
- The standard error of the sample proportion P is (just replace the unknown π appearing in the formula by the known p): $p(1-p)/n$.
- For our example, the standard error of the sample proportion is: $\sqrt{0.43(1-0.43)/79} = 0.0557$.

Inference: Confidence Interval for π

- Does the shape of the histogram remind you of a model?
- Remember, by the CLT, the shape of the distribution of p is normal. Also, remember, we can apply the rule which says that 95% of sample proportions (one of which is ours) will fall within 2 standard deviations (which we estimate by the standard error) of the true mean π.
- Hence, the interval that extends two standard errors from the sample proportion should capture the mean (in 95% of the cases).
- A 95% **confidence interval** for an unknown, true proportion π is given by:

$$p \pm 2\sqrt{p(1-p)/n}$$

Inference: Confidence Interval for π

- In general, $100(1-\alpha)\%$ of sample proportions (one of which is ours) will fall within $z_{1-\alpha/2}$ standard deviations (which we estimate by standard error) of the true mean π, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ percentile of the standard normal distribution.
- A $100(1-\alpha)\%$ confidence interval for an unknown, true proportion π is given by:

$$p \pm z_{1-\alpha/2}\sqrt{p(1-p)/n}$$

- For our Example 2 (using $\alpha = 0.05$):

$$0.43 \pm 1.96\sqrt{0.43(1-0.43)/79} = [0.319, 0.541]$$

- Interpretation: In repeated sampling (that is, where we repeat the experiment several times), in 95 out of 100 cases this interval captures (contains/covers) the true proportion π.

Inference: Confidence Interval for π

- **Interpretation: We are 95% certain that the true proportion of firemen from region 1 is at least 31.9% and at most 54.1%.**
- The part $\pm 2\sqrt{p(1-p)/n}$ is known as the **margin of error**, which you all have seen for results of opinion polls.
- **Caution:** This confidence interval does not work when n or π is small! Rule of thumb: $np > 10$ and $n(1-p) > 10$. (However, when π is really small (e.g., 0.0001), this is also inappropriate and there are easy alternatives.)
- **Question:** The chief fire inspector claims that 30% of firemen are from region 1. Is that a reasonable claim?
- **Answer:** No, since 30% is not contained in the confidence interval for the true proportion of firemen from region 1. It is not a plausible value for the true proportion.

Inference: Hypothesis Testing

- The last question can also be formulated as a **hypothesis test:** The hypothesis is if the true proportion of firemen from region 1 is equal to 30% or different from 30%.
- Question: Is there sufficient evidence from the sample to reject the hypothesis that the true proportion is equal to 30%?
- Since most hypothesis test can be answered via a confidence interval, we cover hypothesis testing only lightly here.
- Every hypothesis test has **4 steps**
- Let's first look at a hypothesis test for a single proportion.

Inference: Hypothesis Test for π

- **Step 1:** Specify null and alternative hypotheses (always about a population parameter):

two-sided:	$H_0 : \pi = \pi_0$	$H_A : \pi \neq \pi_0$
one-sided:	$H_0 : \pi \leq \pi_0$	$H_A : \pi > \pi_0$
one-sided:	$H_0 : \pi \geq \pi_0$	$H_A : \pi < \pi_0,$

  where π **is the true (unknown) proportion** and π_0 is some specific value. Also, choose α-level (controls type I error, see later).

- **Step 2:** Specify the (asymptotic) distribution for the estimator of the unknown parameter. In almost all cases: apply the CLT assuming H_0 is true:

$$P \sim N(\pi_0, \pi_0(1-\pi_0)/n)$$

Inference: Hypothesis Test for π

- **Step 3 (P-value):** Assuming that H_0 is true, find the probability of observing an even more extreme (as specified by the alternative hypothesis) sample proportion as the one observed:

  I.e., in the one sided case with H_A: $\pi > \pi_0$, find $\Pr(P > p)$, where p is the observed proportion.

  In the two-sided case, find $2 \times \Pr(P > |p|)$.

  To find this, calculate the **Test Statistic:** under H_0

$$Z = \frac{P - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}.$$

Inference: Hypothesis Test for π

- Calculate the **Test Statistic:** under H_0

$$Z = \frac{P - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}.$$

The probability $\Pr(P > p)$ is the same as $\Pr(Z > z)$, and $2 \times \Pr(P > |p|)$ is the same as $2 \times \Pr(Z > |z|)$, where Z is a standard normal random variable (i.e., $Z \sim N(0,1)$).

The probability under the $N(0,1)$ model is easy to calculate (Tables, `R`). The resulting probability is known as the P-value.

Inference: Hypothesis Test for π

- **Step 4 (Conclusion):**
  If P-value $< \alpha$: **Sufficient evidence** for H_A.
  - H_0 is no more tenable, reject it. The likelihood of observing such a sample proportion when the null hypothesis is true is so small, so that the null hypothesis must be wrong.

  If P-value $\geq \alpha$: **Insufficient evidence**. Cannot reject the claim H_0, therefore retain it.
  - The sample did not provide overwhelming evidence to reject the null hypothesis. The likelihood of observing such a sample proportion is not so small when the null hypothesis is correct. Therefore, no reason to reject it.
- How to choose α?

Inference: Hypothesis Test for π

- **2 Errors** (wrong decisions):
  - Type I: reject H_0 although H_0 is true;
  - Type II: retain H_0 although H_0 is false (i.e. H_A true)
- The probability of making Type I error equals α:

$$\Pr(\text{reject } H_0 | H_0 \text{ true}) = \alpha.$$

- In hypotheses testing we **control the Type I error** (at $100\alpha\%$) by choosing a small α-level in Step 1 (e.g. 5%, 1%, 10%).
- The Power of a test $= 1-$Type II error = reject H_0 when it is false = reject H_0 when you should!
- The larger the power, the better.
- There is a relationship between Type I error, power and sample size.

Inference: Hypothesis Test for π

- Example 2 (cont.) Test the hypothesis that the true proportion of firemen from region 1 is 30%.
  - **Step 1:** Write out null and alternative hypothesis

$$H_0 : \pi = 0.3 \qquad H_0 : \pi \neq 0.3$$

    and choose $\alpha = 5\%$.
  - **Step 2:** From the CLT we know:

$$P \sim N(0.3, 0.3(1-0.3)/79) = N(0.3, 0.309)$$

  - **Step 3:** Calculate Test statistic

$$z = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = \frac{0.43 - 0.30}{\sqrt{0.3(1-0.3)/79}} = 2.52$$

    and P-value: $2 \times \Pr(Z > |z|) = 2 \times \Pr(Z > 2.52) = 0.0117$.

Inference: Hypothesis Test for π

- This can be easily calculated in R:

```
> 2*(1 - pnorm(2.52))
> [1] 0.01173548
```

- **Step 4:** Conclusion: Since the P-value of 0.0117 is less than the α of 5%, we reject the null hypothesis and conclude that the true proportion of firemen from region 1 must be different from 30%.
- Different by how much?
- Give confidence interval: [31.9%, 54.1%]
- So, **confidence interval more informative** as reject/do not reject decision of statistical hypothesis test.

Inference: Hypothesis Test for π

- There is an `R` function, `prop.test`, that computes confidence intervals and P-values for hypothesis tests of unknown population proportions. However, it uses a slightly different test statistic, so results are slightly different from ours. The procedure we outlined before is generally recommended.

```
> table(region)
 region
  1  2
 34 45
```

Inference: Hypothesis Test for π

- `> prop.test(x=34, n=79, p=0.3)`

```
1-sample proportions test with continuity correction

data: 34 out of 79, null probability 0.3
X-squared = 5.789, df = 1, p-value = 0.01613
alternative hypothesis: true p is not equal to 0.3
95 percent confidence interval:
0.3210938  0.5464879
sample estimates:
        p
0.4303797
```

Inference: Confidence Interval for μ

- We have seen how to construct confidence intervals for a proportion. Now we want to do the same for an unknown mean of a continuous variable in the population.
  Example 1 (cont.): What is a range of plausible values for the true FVC value of firemen?
- Same procedure as before. By the CLT, we know that the standard error of the sample mean $\bar{X}$ is $S/\sqrt{n}$.
- We also know by the CLT that the distribution of the sample mean is normal.
- Hence, by extending 2 standard errors to the left and right of our observed sample mean $\bar{x}$, we should capture the true mean 95% of the time. (In 5% of cases, the sample mean we obtain is so awkward (unlucky), that we will not capture the true sample mean.)

Inference: Confidence Interval for μ

- A 95% confidence interval for an unknown, true mean μ is given by:

$$\bar{x} \pm 2s/\sqrt{n}.$$

- More generally, **a $100(1-\alpha)$% confidence interval for an unknown, true mean μ** is given by

$$\bar{x} \pm z_{1-\alpha/2}s/\sqrt{n}.$$

- Same interpretations as before.

Inference: Confidence Interval for μ

Example 1 (cont.): Find a 95% confidence interval for the mean FVC value in the population of firemen.

```
> x.bar <- mean(fvc)
> s <- sqrt(var(fvc))
> n <- length(fvc))
> alpha <- 0.05
> z <- qnorm(1-alpha/2)
> x.bar + c(+1, -1)*z*s/sqrt(n)
[1] 5.366750  5.703123
```

Inference: Confidence Interval for μ

- We are 95% certain that the true FVC of firemen in the population is at least 5.37 and at most 5.70.
- What does "95% certain" mean: Would we continue to take samples of the same size, in 95 out of 100 cases the interval so constructed contains the true mean. (And we hope that our specific case is one that belongs to the 95 cases).
- Question: A FVC value of 5.4 is considered "normal". Can the average FCV value for firemen be considered normal?
- Answer: Yes! 5.4 lies within the 95% confidence interval, it is therefore a plausible value for the average (mean) FCV value.

Inference: Confidence Interval for μ

- The confidence intervals above are valid if the sample size is "large" (79 is considered large enough).
- For smaller samples, the CLT doesn't work. But we only need one minor adjustment to obtain a valid confidence interval even with small samples.
- However, the price we have to pay is that we have to assume that the random variables which make up our sample are actually iid from a normal distribution $N(\mu, \sigma^2)$.
- For large sample sizes, we didn't need to make this assumption. (The CLT guaranteed normality of the sample mean, no matter what the distribution of the random variables that make up the sample, as long as they are iid.)

Inference: Confidence Interval for μ

- The adjustment we have to make is that instead of the standard normal $N(0,1)$ distribution to calculate the upper $z_{1-\alpha/2}$ quantile (which is often called the critical value), we have to use **Student's t distribution** with $n-1$ degrees of freedom and the corresponding upper $t_{n-1,1-\alpha/2}$ quantile for that distribution (for which there are also Tables, or use R).
- The confidence interval then looks as follows:

$$\bar{x} \pm t_{n-1,1-\alpha/2} s/\sqrt{n}.$$

Inference: Confidence Interval for μ

Example 1 (cont.): Find a 95% CI for mean FVC:

```
> z <- qnorm(1-alpha/2);
> x.bar + c(+1, -1)*z*s/sqrt(n)
[1] 5.366750  5.703123

> t <- qt(1-alpha/2, df=n-1)
> x.bar + c(+1, -1)*t*s/sqrt(n)
[1] 5.364100  5.705773

> c(z, t)
[1] 1.959964  1.990847
```

Inference: Confidence Interval for μ

- The CI is virtually the same. In fact, we almost always use t, because as n gets larger, the critical value based on the t distribution gets closer to the critical value based on the $N(0, 1)$ distribution.
- To answer the question, we can also conduct a hypothesis test for the unknown mean μ.
- Same 4 steps as before, but no with regard to a mean:

Inference: Confidence Interval for μ

- **Step 1 (Setup):** Write out **null** and **alternative** hypotheses (always about a population parameter):

  two-sided: $H_0 : \mu = \mu_0$ vs. $H_A : \mu \neq \mu_0$

  one-sided: $H_0 : \mu = \mu_0$ vs. $H_A : \mu > \mu_0$ (or $H_A : \mu < \mu_0$),

  where μ **is the true (unknown) mean** and μ_0 is some specific value. Also, choose α-level (controls type I error).
- **Step 2 (Specify)**: Specify the (asymptotic) distribution for the estimator of the unknown parameter. In almost all cases: apply the CLT assuming H_0 is true:

$$\bar{X} \sim N(\mu_0, \sigma^2/n).$$

Inference: Confidence Interval for μ

- **Step 3 (Test statistic and P-value):** Assuming that H_0 is true, find the probability of observing an even more extreme (as specified by the alternative hypothesis) sample mean as the one observed: I.e., in the one-sided case with H_A: $\mu > \mu_0$, find $\Pr(\bar{X} > \bar{x})$, where $\bar{x}$ is the sample mean. In the two-sided case, find $2\Pr(\bar{X} > |\bar{x}|)$.
  To find it, calculate **Test Statistic:** $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$.
  The probability $\Pr(\bar{X} > \bar{x})$ is the same as $\Pr(T > t)$, and $2\Pr(\bar{X} > |\bar{x}|)$ is the same as $2\Pr(T > |t|)$, where T is distributed as Student's t with $n - 1$ degrees of freedom. This probability under the t distribution is easy to calculate (Tables, R). The resulting probability is known as the **P-value**, and the entire test is often referred to as the **t-test**.

Inference: Confidence Interval for μ

- **Step 4 (Conclusion):**
  If P-value $< \alpha$: Sufficient evidence for H_A.
  - H_0 is no more tenable, reject it. The likelihood of observing such a sample mean when H_0 were true is too small, so that H_0 must be wrong.
- If P-value $\geq \alpha$: Insufficient evidence. Cannot reject the claim H_0, therefore we retain it.
  - The sample did not provide overwhelming evidence to reject H_0. The likelihood of observing such a sample mean is not so small when H_0 is correct. Therefore, no reason to reject it.

Inference: Confidence Interval for μ

Example 1 (cont.) Is there reason to believe that firemen have an abnormal FCV? (I.e., is there mean FCV different from the normal value of 5.4?)

- Step 1: Write out null and alternative hypothesis

$$H_0 : \mu = 5.4 \quad \text{vs.} \quad H_A : \mu \neq 5.4$$

  and set $\alpha = 5\%$.

- Step 2: From the CLT we know: $\bar{X} \sim N(5.4, \sigma^2/n)$.
- Step 3: Calculate test statistic

$$t = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{5.535 - 5.4}{0.763/\sqrt{79}} = 1.5725$$

  and P-value: $2\Pr(T > |t|) = 2\Pr(T > 1.57) = 0.1199$.

Inference: Confidence Interval for μ

- **Conclusion: We have insufficient evidence (P = 0.1199) to conclude that firemen have an abnormal mean FVC.**
- We can get all results (CI and test) with one function in R:

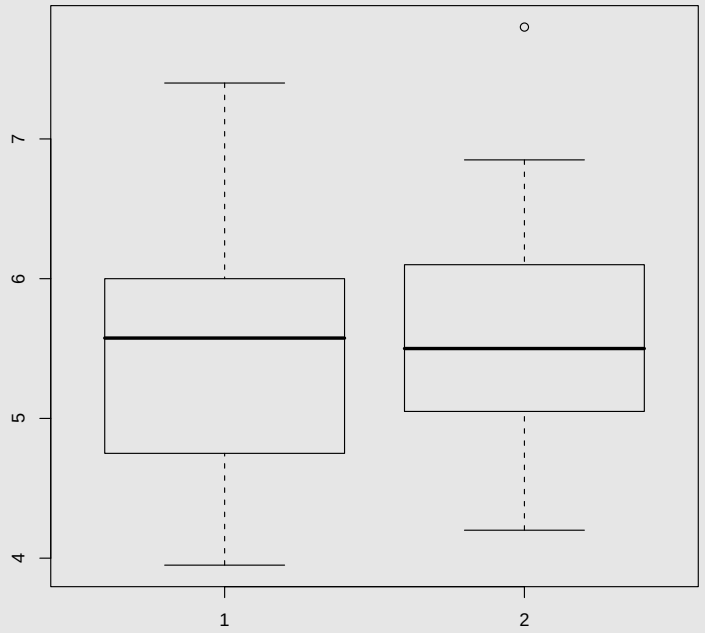
```
> t.test(fvc, mu=5.4)
        One Sample t-test
 data: fvc
 t = 1.5725, df = 78, p-value = 0.1199
 alternative hypothesis: true mean is not equal to 5.4
 95 percent confidence interval:
 5.364100 5.705773
 sample estimates:
 mean of x
 5.534937
```

Inference: 2 Samples

- Often, we are interested in comparing population parameters from two groups.
- Example 2 (cont.): Is there a difference between the average FVC from firemen in region 1 and 2?
- First, create a plot to compare the two samples graphically: Two Boxplots, side by side:

```
> boxplot(fvc ~ region)
```

Inference: 2 Samples



It seems that there are no great differences. Let's find out more precisely, by finding a range of plausible values for the difference in mean FVC between firemen from region 1 and 2.

Inference: CI for $\mu_1 - \mu_2$

- We have a random sample of n_1 observations from population 1, and a random sample of n_2 observations from population 2.
- To construct CI, basically same procedure as before. Appeal to CLT:

$$\begin{aligned} & \bar{X}_1 \sim N(\mu_1, \sigma_1^2/n_1)\,, \quad \bar{X}_2 \sim N(\mu_2, \sigma_2^2/n_2) \\ \Longrightarrow \quad & \bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2). \end{aligned}$$

- Hence, the standard error for the differences in sample means $\bar{X}_1 - \bar{X}_2$ is given by:

$$\sqrt{S_1^2/n_1 + S_2^2/n_2}.$$

Inference: CI for $\mu_1 - \mu_2$

- Remember, once we have standard error (i.e., a measure how variable our estimate of the true differences is), we can construct confidence intervals.
- The confidence interval for the difference between two population means $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t_{df,1-\alpha/2}\sqrt{s_1^2/n_1 + s_2^2/n_2}$$

  where $t_{df,1-\alpha/2}$ is the upper $\alpha/2$ quantile of the t distribution with degrees of freedom df (a complicated formula, use software).
- Lets again use the R function `t.test` to get this confidence interval.

Inference: CI for $\mu_1 - \mu_2$

```
> t.test(fvc ~ region)
 Welch Two Sample t-test
 data: fvc by region
 t = -0.8696, df = 65.771, p-value = 0.3877
 alternative hypothesis: true difference in means is not equal t
 95 percent confidence interval:
 {-0.5067853 0.1992690
 sample estimates:
 mean in group 1 mean in group 2
 5.447353 5.601111
```

- Interpretation: We are 95% confident that the difference between the mean FCV of firemen from region 1 and the mean FCV of firemen from region 2 is between -0.51 and 0.20. Since 0 is contained in this interval, there is no significant difference between the mean FCV for these two groups.

Inference: CI for $\mu_1 - \mu_2$

```
> t.test(fvc ~ region)
 Welch Two Sample t-test
 data: fvc by region
 t = -0.8696, df = 65.771, p-value = 0.3877
```

- The t-value, df and P-value refer to a hypothesis test for the difference between two means $\mu_1 - \mu_2$:
- **Test (two independent sample t-test)**

$$H_0 : \mu_1 - \mu_2 = 0 \text{ vs. } H_A : \mu_1 - \mu_2 \neq 0 \quad (\text{or } <, >)$$

- **Test Statistic**

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

refers to t_{df} distribution.

Inference: CI for $\mu_1 - \mu_2$

- Interpretation of P-value: Since the P-value is large in comparison to any reasonable value for α (e.g., 1%, 5% or even 10%), we have insufficient evidence to reject the null hypothesis that the two mean FVC values are different.
- Again: Hypothesis test and confidence interval give same conclusion, but confidence interval provides more information about the size of the difference.
- Finally, similar to finding confidence intervals for difference of means, we can ask for differences of proportions:

Example 3: Is the proportion of firemen older than 20 years the same in region 1 and 2?

Inference: CI for $\mu_1 - \mu_2$

- To answer this question, we construct a confident interval for the difference between the true (but unknown) proportions of firemen older than 20 years in region 1 and 2.
- What are the sample proportions in the two groups?

```
> age20 <- (age > 20)
> table(region, age20)
            age20
 region  FALSE  TRUE
      1     15    19
      2      6    39
```

- So, 19/(15+19) = 55.9% of firemen in region 1 are older than 20 years, while 39/(6+39) = 86.7% of firemen in region 2 are older than 20 years. Is this observed difference of 30.8% significant, or just due to sampling variability?

Inference: CI for $\pi_1 - \pi_2$

- We have a random sample of n_1 observations from population 1 (yielding the first sample proportion p_1), and a random sample of n_2 observations from population 2 (yielding the second sample proportion p_2)
- To construct a CI, basically same procedure as before. Appeal to CLT:

$$P_1 \sim N(\pi_1, \pi_1(1-\pi_1)/n_1), \quad P_2 \sim N(\pi_2, \pi_2(1-\pi_2)/n_2)$$
$$\Longrightarrow \quad P_1 - P_2 \sim N(\pi_1 - \pi_2, \pi_1(1-\pi_1)/n_1 + \pi_2(1-\pi_2)/n_2).$$

- Hence, the standard error for the differences in sample proportions $P_1 - P_2$ is given by:

$$\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}.$$

Inference: CI for $\pi_1 - \pi_2$

- Remember, once we have standard error (i.e., a measure how variable our estimate of the true differences is), we can construct CIs.
- The confidence interval for the difference between two population proportions $\pi_1 - \pi_2$ is

$$(p_1 - p_2) \pm z_{1-\alpha/2} \sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2},$$

  where $z_{1-\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution $N(0,1)$.
- The R function `prop.test` delivers this interval.

Inference: CI for $\pi_1 - \pi_2$

```
> table(region, age20)
            age20
 region  FALSE  TRUE
      1     15    19
      2      6    39
> prop.test(x=c(19,39), n=c(34,45))
    2-sample test for equality of proportions
    with continuity correction
  data: c(19, 39) out of c(34, 45)
  X-squared = 7.8931, df = 1, p-value = 0.004962
  alternative hypothesis: two.sided
  95 percent confidence interval:
   -0.5278754 -0.0878109
  sample estimates:
   prop 1 prop 2
  0.5588235 0.8666667
```

Inference: CI for $\pi_1 - \pi_2$

- Interpretation: The CI for the true difference in proportions is [-53%, -9%]. That is, the true proportion of firemen older than 20 years is lower in region 1 when compared to region 2 by at least 9% and at most 53%.
- We can also say: The true proportion of firemen older than 20 years is at least 9% and at most 53% larger in region 2.
- The R output also displays results (i.e., the P-value) of a hypothesis test for the hypothesis that the true difference is zero:

$$H_0 : \pi_1 - \pi_2 = 0 \qquad H_A : \pi_1 - \pi_2 \neq 0.$$

Inference: CI for $\pi_1 - \pi_2$

- Conclusion: Since the P-value is very small (compared to an α of 5% or even 1%), we have sufficient evidence to reject the null hypothesis that the difference in true proportions is zero (i.e., that the true proportions are equal). Therefore, the two proportions are significantly different.
- A CI gives more information then the hypothesis test. It also tells the size of the effect, i.e., by how much the true proportions differ.

Summary

- To describe distribution of a variable: shape, center, spread.
- Typical parameters to describe the center are the mean (for continuous variables) and the proportion (for binary variables).
- We learned how to estimate these by the sample mean and sample proportion, respectively.
- We learned that these are reasonable estimates: they are unbiased.
- We learned how to assess the variability (=precision) of these estimators by finding their standard error.
- The estimate plus the standard error combine to give a CI for the true population parameter.
- We learned how to conduct hypothesis tests about these parameters.
- Finally, we looked at comparing the mean and proportion among two groups via constructing a CI for their difference.