Hypothesis Testing

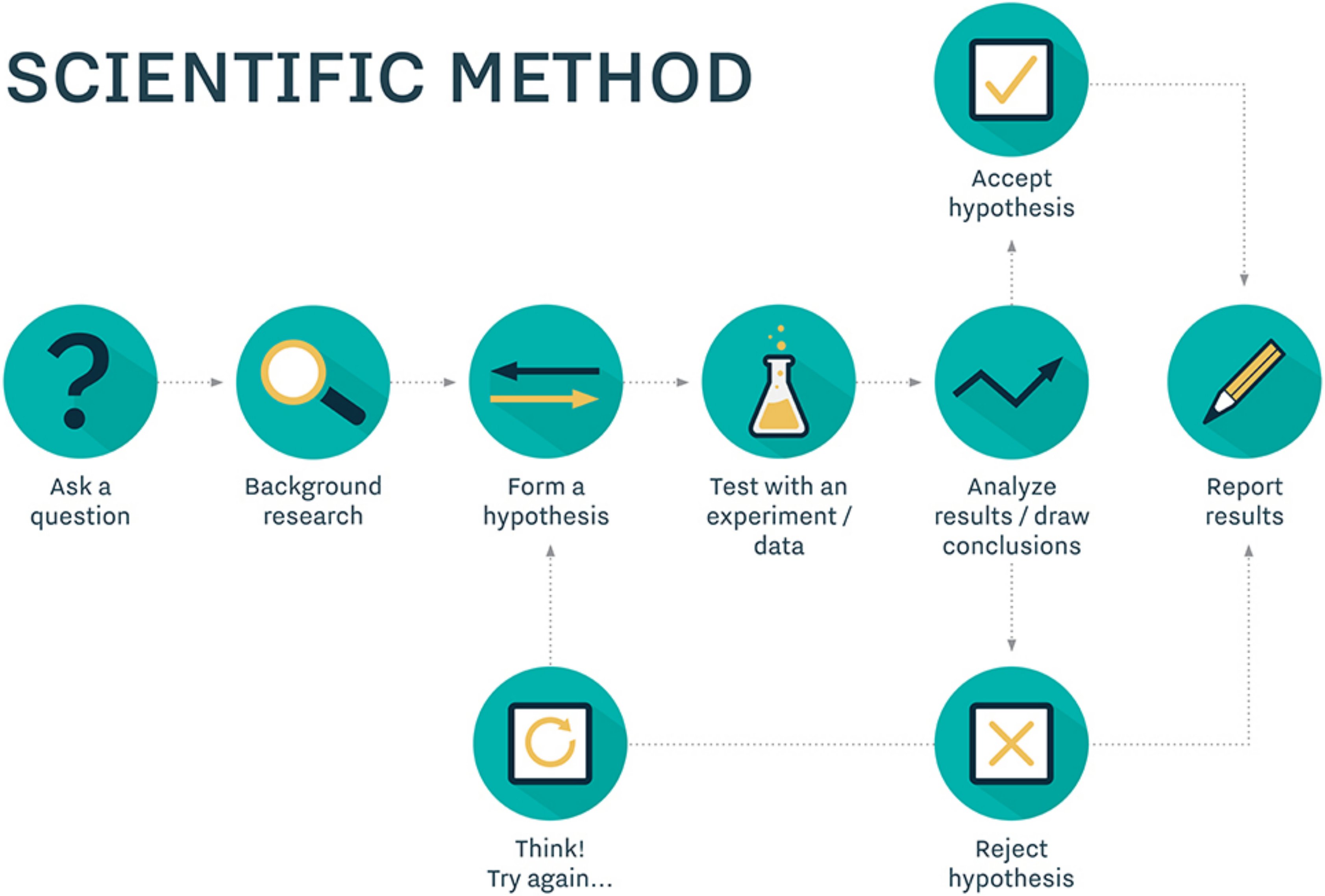
Hypothesis

- A hypothesis (*plural: hypotheses*) is a testable statement about the relationship between two or more variables or a proposed explanation for some observed phenomenon.
- In a scientific experiment or study, the hypothesis is a brief summation of the researcher's prediction of the study's findings, which may be supported or not supported by the outcome.
- Hypothesis testing is the core of the scientific method.

Hypothesis

- Scientific method is an approach to seeking knowledge that involves forming and testing a hypothesis.
- Scientific method provides a logical, systematic way to answer questions and removes subjectivity by requiring each answer to be authenticated with objective evidence that can be reproduced.
- Goal of scientific method is to gather data that will validate or invalidate a cause and effect relationship.
- Scientific method is often carried out in a linear manner, but the approach can also be cyclical, because once a conclusion has been reached, it often raises more questions.

Hypothesis


SCIENTIFIC METHOD
Accept hypothesis
Ask a question
Background research
Form a hypothesis
Test with an experiment / data
Analyze results / draw conclusions
Report results
Think! Try again...
Reject hypothesis

Hypothesis

- In general, many scientific investigations start by expressing a hypothesis.
- To evaluate hypotheses, we rely on
  - estimators,
  - their sampling distributions,
  - their specific values

  from observed data.
- For example,
  - Mackowiak et al.* hypothesized that the average normal (i.e., for healthy people) body temperature is less than the widely accepted value of 98.6°F.
  - If we denote the population mean of normal body temperature as μ, then we can express this hypothesis as $\mu < 98.6$.

*Mackowiak, P.A., Wasserman, S.S., Levine, M.M.: A critical appraisal of 98.6°F, the upper limit of the normal body temperature, and other legacies of Carl Reinhold AugustWunderlich. JAMA 268, 1578–1580 (1992)

Null and Alternative hypotheses

- The null hypothesis usually reflects the “status quo" or “nothing of interest".
- In contrast, we refer to our hypothesis (i.e., the hypothesis we are investigating through a scientific study) as the alternative hypothesis and denote it as H_A.
- Consider the body temperature example, where we want to examine the null hypothesis $H_0 : \mu = 98.6$ against the alternative hypothesis, $H_A : \mu < 98.6$.
- The procedure for evaluating a hypothesis is called hypothesis testing, and it rises in many scientific problems.
- **For hypothesis testing, we focus on the null hypothesis since it tends to be simpler.**
- To this end, we examine the evidence that the observed data provide against the null hypothesis H_0.
  - If the evidence against H_0 is strong, we reject H_0.
  - If not, we state that the evidence provided by the data is not strong enough to reject H_0, and we fail to reject it.

Null and Alternative hypotheses

- With respect to our decision regarding the null hypothesis H_0, we might make two types of errors:
  - Type I error:
    - we reject H_0 when it is true and should not be rejected.
  - Type II error:
    - we fail to reject H_0 when it is false and should be rejected.
- We denote the probability of making type I error as $\boldsymbol{\alpha}$ and the probability of making type II error as $\boldsymbol{\beta}$.

Image source: unbiasedresearch.blogspot.com

	Actual Validity of H_0		
Decision Made		H_0 is true	H_0 is false
	Accept H_0	True Negative	False Negative (Type II Error)
	Reject H_0	False Positive (Type I Error)	True Positive

Null and alternative hypotheses

- Now suppose that we have a hypothesis testing procedure that fails to reject the null hypothesis when it should be rejected with probability β.
  - This means that our test correctly rejects the null hypothesis with probability $1 - \beta$.
    - Note that the two events are complementary.
  - We refer to this probability (i.e., $1 - \beta$) as the power of the test.
- In practice, it is common to first agree on a tolerable type I error rate α, such as 0.01, 0.05, and 0.1.
- Then try to find a test procedure with the highest power among all reasonable testing procedures.

Hypothesis testing for the population mean

- To decide whether we should reject the null hypothesis, we quantify the empirical support (provided by the observed data) against the null hypothesis using some statistics.
- We use statistics to evaluate our hypotheses.
  - We refer to them as test statistics.
    - To evaluate hypotheses regarding the population mean, we use the sample mean $\bar{X}$ as the test statistic
- For a statistic to be considered as a test statistic, its sampling distribution must be fully known (exactly or approximately) under the null hypothesis.
  - We refer to the distribution of test statistics under the null hypothesis as the null distribution.
    - For the sample mean, the CLT states that the sampling distribution is approximately normal when the sample size is large.

Hypothesis testing for the population mean

- Consider the body temperature example, where we want to examine the null hypothesis $H_0 : \mu = 98.6$ against the alternative hypothesis $H_\text{A} : \mu < 98.6$.
- To start with, suppose that $\sigma^2 = 1$ is known and we have randomly selected a sample of 25 healthy people from the population and measured their body temperature.
- Using the CLT, the sampling distribution of $\bar{X}$ is approximately normal as follows:

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

- For the above example,

$$\bar{X} \sim N(\mu, 1/25)$$

- If the null hypothesis is true and the population mean is $\mu = 98.6$,
- the sampling distribution of $\bar{X}$ becomes

$$\bar{X} | H_0 \sim N(98.6,\ 0.04)$$

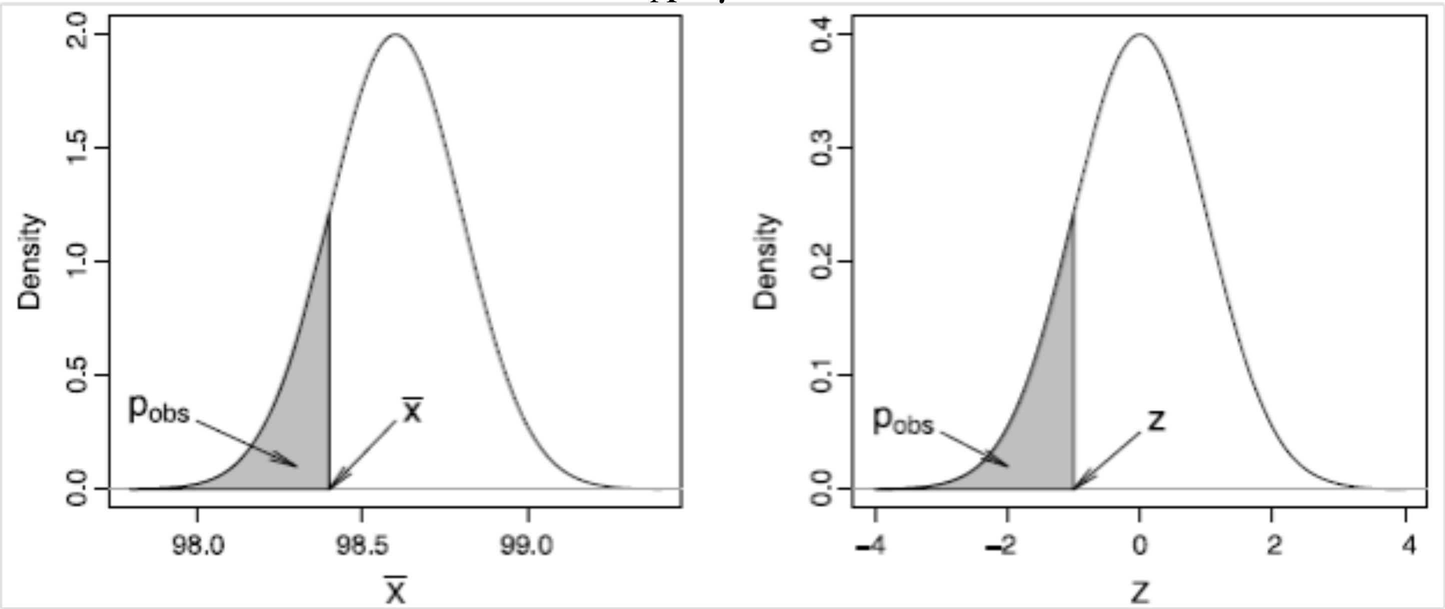
  - Note that the distribution of $\bar{X}$ is obtained conditional (hence the notation for conditional probability) on the assumption that the null hypothesis is true

Hypothesis testing for the population mean

- In reality, we have one value, $\overline{x}$, for the sample mean.
- We can use this value to quantify the evidence of departure from the null hypothesis.
- Suppose that from our sample of 25 people we find that the sample mean is $\overline{x} = 98.4$
- To evaluate the null hypothesis $H_0 : \mu = 98.6$ versus the alternative $H_A : \mu < 98.6$,
  - we use the lower tail probability of this value from the null distribution.

Hypothesis testing for the population mean

- For the normal body temperature example, examining the hypotheses $H_0 : \mu = 98.6$ versus the alternative $H_A : \mu < 98.6$.



- *Left panel*: The *shaded area* shows the lower-tail probability of the observed sample mean, $\bar{x} = 98.4$. This is the observed significance level, *p*-value, which is denoted as p_{obs}.
- *Right panel*: After standardizing, the *p*-value corresponds to the lower tail probability of $z = -1$ based on the standard normal distribution

Observed significance level

- The observed significance level for a test is the probability of values as or more extreme than the observed value, based on the null distribution in the direction supporting the alternative hypothesis.
- This probability is also called the p-value and denoted as p_{obs}.
- For the above example,

$$p_{obs} = P\left(\overline{X} \leq \overline{x} \middle| H_0\right) = P(\overline{X} \leq 98.4) = 0.16$$

- To find the *p*-value in R-Commander,
  - click *Distributions* → *Continuous distributions* → *Normal distribution* → *Normal probabilities*.
  - Then set the *Variable value* to 98.4 and the parameters for the null distribution ($\mu = 98.6$ and $\sigma = 0.2$).

z-Tests of the Population Mean

- In practice, it is more common to use the standardized version of the sample mean as our test statistic.
- We know that if a random variable is normally distributed (as it is the case for $\overline{X}$), subtracting the mean and dividing by standard deviation creates a new random variable with standard normal distribution, $Z \sim N(0,1)$
- We refer to the standardized value of the observed test statistic as the z-score,

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{98.4 - 98.6}{0.2} = -1$$

$$p_{obs} = P(Z \leq -1) = 0.16$$

- We refer to the corresponding hypothesis test of the population mean as the z-test.
- In a z-test, instead of comparing the observed sample mean $\bar{x}$ to the population mean according to the null hypothesis, we compare the z-score to 0

Interpretation of p-value

- The p-value is the conditional probability of extreme values (as or more extreme than what has been observed) of the test statistic assuming that the null hypothesis is true.
  - When the p-value is small, say 0.01 for example, it is rare to find values as extreme as what we have observed (or more so).
- As the p-value increases, it indicates that there is a good chance to find more extreme values (for the test statistic) than what has been observed.
  - Then, we would be more reluctant to reject the null hypothesis.
- The common significance levels are 0.01, 0.05, and 0.1.

Interpretation of p-value

- If p_{obs} is less than the assumed cutoff, we say that the data provides statistically significant evidence against H_0
- For the body temperature example where $p_{\text{obs}} = 0.16$, if we set the significance level at 0.05, we say that there is not significant evidence against the null hypothesis H_0: μ = 98.6 at the 0.05 significance level, so we do not reject the null hypothesis.
- If we had set the cutoff at 0.05 and we had observed $\overline{x} = 98.25$ instead of 98.4, then $p_{\text{obs}} = 0.04$, and we could reject the null hypothesis.
- In this case, we say that the result is statistically significant and the data provide enough evidence against $H_0 : \mu =$ 98.6.

One-Sided Hypothesis Testing

- In general, for one-sided hypothesis testing, we evaluate the null hypothesis $H_0 : \mu = \mu_0$ by using the following standardized test statistic:

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

- We find the sample mean $\overline{x}$ and calculate the observed value of Z called z-score (assuming σ is known):

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

One-Sided Hypothesis Testing

- We then use the standard normal distribution to find the *p*-value.
- If the alternative hypothesis regarding the population mean is $H_A : \mu < \mu_0$, we use the standard normal distribution to find lower tail probability of the *z*-score: $P(Z \leq z)$.
- If the alternative hypothesis regarding the population mean is $H_A : \mu > \mu_0$, we use $P(Z \geq z)$ instead (upper tail probability).
- The resulting probability, p_{obs}, is the observed significance level, which can be compared to several significance levels such as 0.01, 0.05, and 0.1.

One-Sided Hypothesis Testing

- In some situations, we might hypothesize that the population mean is greater than a specific value and express our hypothesis as $H_A : \mu > \mu_0$.
- Our null hypothesis is still $H_0 : \mu = \mu_0$.
- This is also a one-sided test since the departure from the null is still in one direction: toward values larger than μ_0.
- For example, suppose that we have observed that many Pima Indian women suffer from diabetes.
- We know that obesity and diabetes are related; we might therefore hypothesize that this population is obese on average, where obesity is defined as BMI higher than 30.
- If we denote the population mean of BMI for Pima Indian women, we can then express our hypothesis as $\mu > 30$.
- In this case, the null hypothesis is $H_0 : \mu = 30$; that is, $\mu_0 = 30$.

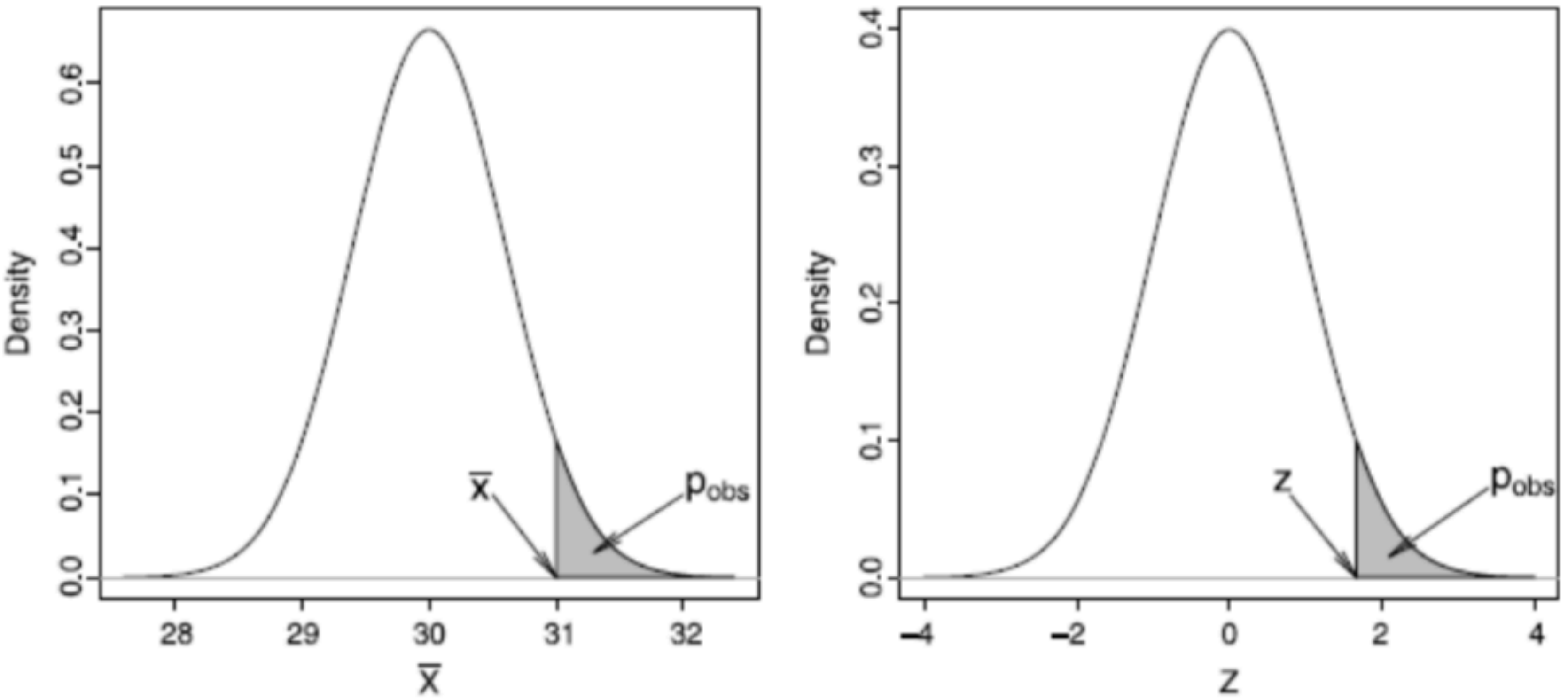
One-Sided Hypothesis Testing

- As before, we use the sample mean as the test statistic. For illustrative purposes,
- suppose that we have obtained a sample of size n = 100 from the population of Pima
- Indian women. Further, suppose we know that the population variance is $\sigma^2 = 6^2$.
- If the null hypothesis is true and the population mean is $\mu = 30$, then the sampling distribution is

$$\bar{X}|H_0 \sim N(30, 6^2/100)$$

- The distribution is shown in the left panel of the figure in the next slide

One-Sided Hypothesis Testing



- *Left panel*: The sampling distribution for the test statistic under the null hypothesis $H_0 : \mu = 30$. The *p*-value, which is the probability of values as or more extreme than the observed value of the test statistic $\bar{x} = 31$, is shown as the *shaded area*.
- *Right panel*: Obtaining the upper tail probability using one-sided *z*-test

One-Sided Hypothesis Testing

- If the null hypothesis is indeed true, then we would expect to see the value of sample mean near the population mean according to the null distribution (here, 30).
- In contrast, if the null hypothesis is false, then the null distribution does not represent the sampling distribution of the test statistics, and we would expect to see the value of the sample mean away from 30, in this case, larger than 30 according to the alternative hypothesis.
- Suppose that from our sample of 100 Pima Indian women we find that the sample mean is $\overline{x} = 31$.
- As before, we find the observed significance level, p-value, to measure the amount of evidence provided by the data in support for H_0.

One-Sided Hypothesis Testing

- Recall that we defined p-value as the probability of values as or more extreme than the observed value of the test statistic (here, $\overline{x} = 31$) based on the null distribution, in the direction specified by the alternative hypothesis.
- If the null distribution is in fact true and $\mu = 30$, then values larger than $\overline{x} = 31$ would seem more extreme than what we have observed.
- Therefore, since $H_A : \mu > \mu_0$

$$p_{\text{obs}} = P(\overline{X} \geq \overline{x} | H_0)$$

- If we drop H_0 for simplicity,

$$p_{\text{obs}} = P(\overline{X} \geq 31)$$

- This probabilitv is shown as the shaded area in the left panel of the figure (in slide 21)

One-Sided Hypothesis Testing

- We can standardize the test statistic by subtracting the mean and dividing the result by the standard deviation:

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X} - 30}{\frac{6}{\sqrt{100}}} = \frac{\overline{X} - 30}{0.6} \sim N(0, 1)$$

- The corresponding *z*-score is obtained as follows:

$$z = \frac{\overline{x} - \mu_0}{0.6} = \frac{31 - 30}{0.6} = 1.67$$

- Now, to find the *p*-value, we can find the upper tail probability of $z = 1.67$ from the null distribution $N(0, 1)$:

$$p_{\text{obs}} = P(Z \geq 1.67)$$

- This probability is shown as the shaded area in the right panel of the figure (in slide 21)
    - This is the upper tail probability at 1.67 based on the standard normal distribution.

One-Sided Hypothesis Testing

- The upper tail probability is by convention $P(Z > 1.67)$.
- However, for continuous random variables, $P(Z > 1.67) = P(Z \geq 1.67)$ since the probability of any specific value (here, 1.67) is 0.
- For this example, $p_{\text{obs}} = 0.048$.
- We can reject the null hypothesis at 0.05 level but not at 0.01 level.
- At 0.05 level, we can conclude that the population mean of BMI for Pima Indian women is higher than 30 and the difference is statistically significant.

Two-Sided Hypothesis Testing

- For many hypothesis testing problems, we might be indifferent to the direction of departure from the null value.
  - In such cases, we can express the null and alternative hypotheses as $H_0 : \mu = \mu_0$ and $H_A : \mu \neq \mu_0$, respectively.
  - Then we consider both large positive values and small negative values of z-score as evidence against the null hypothesis, and our alternative hypothesis is referred to as two-sided.
- The p-value for the two-sided hypothesis test is calculated as follows (assuming σ is known):
  - Determine the observed z-score: $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
  - Take the absolute value of the z score: $|z|$.
  - Obtain the upper tail probability: $P(Z \geq |z|)$.
  - Double the resulting probability: $p_{\text{obs}} = 2 \times P(Z \geq |z|)$.

Two-Sided Hypothesis Testing

- For example, suppose we believe that the average normal body temperature is different from the accepted value 98.6°F, but we are not sure whether it is higher or lower than 98.6.
- Then the null hypothesis remains $H_0 : \mu = 98.6$, but the alternative hypothesis is expressed as $H_{\mathrm{A}} : \mu \neq 98.6$.
- We calculate the sample mean $\bar{x} = 98.4$ and standardize it to obtain the z-score,

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{98.4 - 98.6}{1/\sqrt{25}} = \frac{98.4 - 98.6}{0.2} = \frac{-0.2}{0.2} = -1$$

Two-Sided Hypothesis Testing

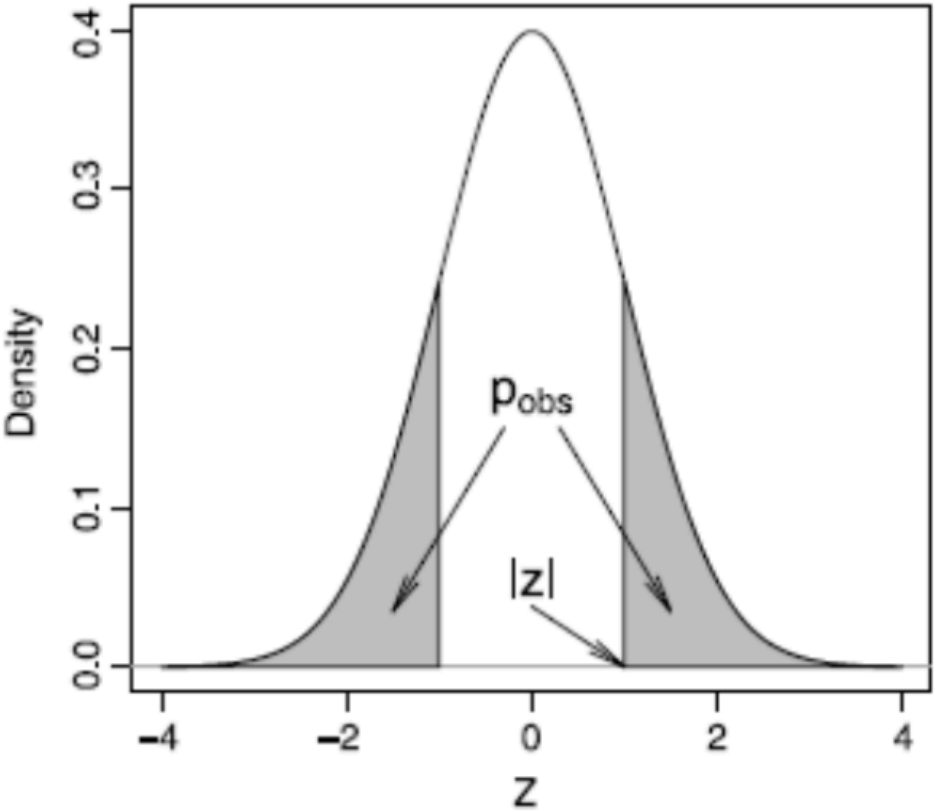
- The *p*-value is still calculated as the probability of values as or more extreme than the observed *z*-score.
  - However, in this case, extreme values are those whose distance from 0 is more than the distance of -1 from zero.
  - These are values that are either less than -1 or greater than 1.
- Therefore, to find the observed significance level, we need to add the probabilities for $Z \le -1$ and $Z \ge 1$:

$$p_{\text{obs}} = P(\mathrm{Z} \le \text{-}1) + P(\mathrm{Z} \ge 1)$$

- This probability is equal to the shaded area in the following figure

Two-Sided Hypothesis Testing

- Illustrating the *p*-value for a two-sided hypothesis test of average normal body temperature, where $H_0 : \mu = 98.6$ and $H_A : \mu \neq 98.6$.



- After standardizing,

$$p_{\text{obs}} = P(Z \leq -1) + P(Z \geq 1) = 2 \times 0.16 = 0.32$$

- The p-value is greater than typical significance levels such as 0.01, 0.05, and 0.1, so we cannot reject the null hypothesis at these levels.
- Therefore, we conclude that the observed difference is not statistically significant, and could be due to chance alone.

Hypothesis testing using t-tests

- So far, we have assumed that the population variance σ^2 is known
  - In reality, σ^2 is almost always unknown, and we need to estimate it from the data.
- As before, we estimate σ^2 using the sample variance S^2.
- Similar to our approach for finding confidence intervals, we account for this additional source of uncertainty by using the t-distribution with n-1 degrees of freedom instead of the standard normal distribution.
  - The hypothesis testing procedure is then called the t-test.

Hypothesis testing using t-tests

- To perform a *t*-test , we use the following test statistic:

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

- The test statistic, T , has a *t*-distribution with n −1 degrees of freedom under the null.

$$T \sim t(n-1)$$

- Using the observed values of $\bar{X}$ and S, the observed value of the test statistic is obtained as follows:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- We refer to *t* as the *t*-score.

Hypothesis testing using t-tests

- To assess the null hypothesis $H_0 : \mu = \mu_0$ using the t -test,
  - we first calculate the t-score based on the observed sample mean $\bar{x}$ and sample standard deviation.
- We then calculate the corresponding p-value as follows:
  - if $H_A : \mu < \mu_0, \quad p_{\text{obs}} = P(T \leq t),$
  - if $H_A : \mu > \mu_0, \quad p_{\text{obs}} = P(T \geq t),$
  - if $H_A : \mu \neq \mu_0, \quad p_{\text{obs}} = 2 \times P(T \geq |t|),$

  where T has a t-distribution with $n - 1$ degrees of freedom, and t is our observed t-score.
- This is known as the single-sample t-test.

Hypothesis testing using t-tests - example

- Suppose we hypothesize that the population mean of BMI among Pima Indian women is above 30: $H_{\mathrm{A}} : \mu > 30$.
- The corresponding null hypothesis is $H_0 : \mu = 30$.
- To test this hypothesis, we use the Pima.tr data set from the MASS package.
- The sample size is $n = 200$.
- The sample mean and standard deviation are $\overline{x} = 32.31$ and $s = 6.13$, respectively. The *t*-score is

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{32.31 - 30}{6.13/\sqrt{200}} = 5.33$$

Hypothesis testing using t-tests - example

- For the above example, $p_{\text{obs}} = P(T \geq 5.33)$, which we obtain from the *t*-distribution with $200 - 1 = 199$ degrees of freedom.
- To obtain this probability in R-Commander,
  - click *Distributions* → *Continuous Distributions* → *t distribution* → *t probabilities*.
  - Then enter 5.33 for *Variable value* and 199 for *Degrees of freedom*, and select *Upper tail*.
  - The resulting probability is 1.33×10^{-07}.
- At any reasonable significance level, there is strong evidence to reject the null hypothesis and conclude that the population mean of BMI among Pima Indian women is in fact greater than 30.
  - Therefore, on average, the population is obese.

Hypothesis testing for population proportion

- For a binary random variable *X* with possible values 0 and 1, we are typically interested in evaluating hypotheses regarding the population proportion of the outcome of interest, denoted as $X = 1$.
- As discussed before, the population proportion is the same as the population mean for such binary variables.
- So we follow the same procedure as described above.
- More specifically, we use the *z*-test for hypothesis testing.
- Note that we do not use *t*-test, because for binary random variable, population variance is $\sigma^2 = \mu(1 - \mu)$.

Hypothesis testing for population proportion

- In general, to assess the null hypothesis $H_0 : \mu = \mu_0$, where μ is the population proportion (mean) of a binary random variable,
  - we first calculate z-score based on the observed sample proportion p:

$$z = \frac{p - \mu_0}{\sqrt{\mu_0 (1 - \mu_0)/n}}$$

- Then we determine the support for the null hypothesis as:
  - if $H_A : \mu < \mu_0$, $p_{\text{obs}} = P(Z \le z)$,
  - if $H_A : \mu > \mu_0$, $p_{\text{obs}} = P(Z \ge z)$,
  - if $H_A : \mu \ne \mu_0$, $p_{\text{obs}} = 2 \times P(Z \ge |z|)$,

  where Z has the standard normal distribution, and z is the observed z-score.

Hypothesis testing for population proportion

- Consider the Melanoma dataset available from the MASS package.
- Suppose that we hypothesize that less than 50% of cases ulcerate, thus: $\mu < 0.5$.
- Then the null hypothesis can be expressed as $H_0 : \mu = 0.5$.
- Using the Melanoma data set, we can test the above null hypothesis
- The number of observations in this data set is $n = 205$, of which 90 patients had ulceration.
- Therefore,

$$p = 90/205 = 0.44$$

Hypothesis testing for population proportion

- Next, we can find the *z*-score for our test statistic as follows:

$$z = \frac{p - \mu_0}{\sqrt{\mu_0\,(1 - \mu_0)/n}} = \frac{0.44 - 0.5}{\sqrt{0.5\,(1 - 0.5)/205}} = -1.72$$

- Because $H_\mathrm{A} : \mu < 0.5$, the observed significance level based on this *z*-score is the lower tail probability $P(Z \leq -1.72)$.
- Using R-Commander, we find the *p*-value to be $p_\mathrm{obs} = 0.043$.
- Therefore, we can reject the null hypothesis at 0.05 level but not at 0.01 level.

Test of Normality

- Visual inspection of the distribution may be used for assessing normality, although this approach is usually unreliable and does not guarantee that the distribution is normal.
  - The frequency distribution (histogram),
  - stem-and-leaf plot,
  - boxplot,
  - P-P plot (probability-probability plot),
  - Q-Q plot (quantile-quantile plot)

  are used for checking normality visually

Test of Normality

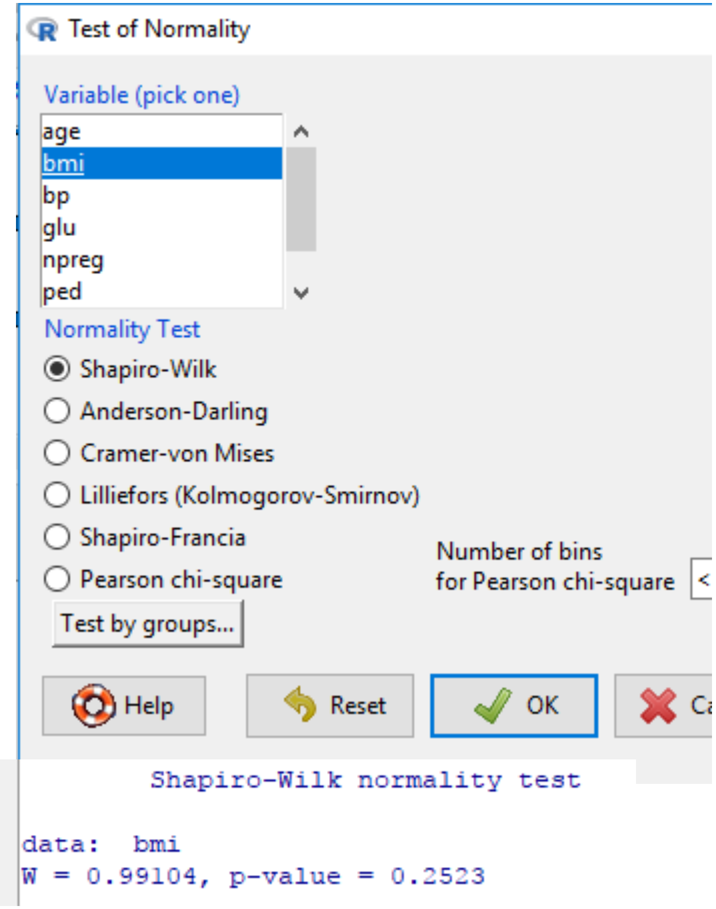
- The appropriateness of the normality assumption can be evaluated formally using a testing procedure such as
  - Kolmogorov-Smirnov (K-S) test
  - Lilliefors corrected K-S test
  - Shapiro-Wilk test
  - Anderson-Darling test
  - Cramer-von Mises test
  - D'Agostino skewness test
  - Anscombe-Glynn kurtosis test
  - D'Agostino-Pearson omnibus test
  - the Jarque-Bera test

Test of Normality - Example

- More specifically, this test evaluates the null hypothesis that the distribution of the random variable is normal.
- As usual, we then either reject this hypothesis and conclude that the normality assumption is not appropriate, or fail to reject it and conclude that there is no strong evidence of deviation from normality.
- Suppose we assume that the *bmi* variable in Pima.tr has a normal distribution.
- To evaluate this assumption, In R-Commander,
  - click *Statistics* → *Summaries* → *Test of normality* and chose *Shapiro-Wilk*,
  - then select the *bmi*.
- The *p*-value for this test is 0.25.
- Therefore, we do not reject the null hypothesis (which states that the distribution is normal) and conclude that the deviation of the distribution from normality is not statistically significant.

Test of Normality - Example

- For comparison, repeat the above steps to test the normality assumption for the *age* variable in the Pima.tr data set.
- Using the Shapiro–Wilk test, the p-value is 1.853×10^{-12}, which is very small.
- Therefore, we can comfortably reject the null hypothesis and conclude that the deviation from normality is statistically significant.



Test of Normality

Variable (pick one)

age
bmi
bp
glu
npreg
ped

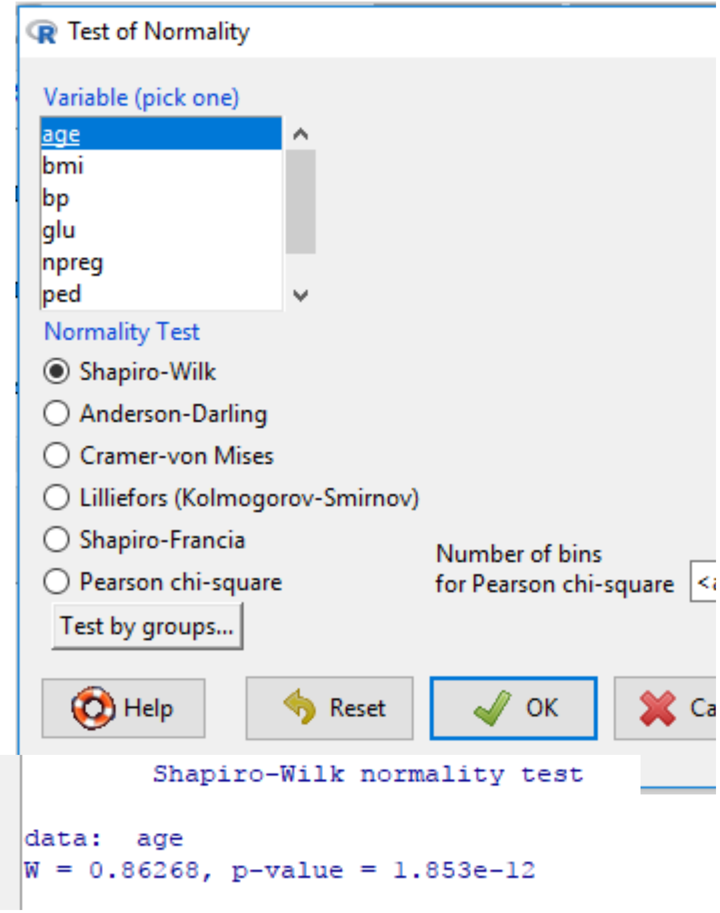
Normality Test

● Shapiro-Wilk
○ Anderson-Darling
○ Cramer-von Mises
○ Lilliefors (Kolmogorov-Smirnov)
○ Shapiro-Francia
○ Pearson chi-square

Number of bins for Pearson chi-square

```
        Shapiro-Wilk normality test

data:  bmi
W = 0.99104, p-value = 0.2523
```



Test of Normality

Variable (pick one)

age
bmi
bp
glu
npreg
ped

Normality Test

● Shapiro-Wilk
○ Anderson-Darling
○ Cramer-von Mises
○ Lilliefors (Kolmogorov-Smirnov)
○ Shapiro-Francia
○ Pearson chi-square

Number of bins for Pearson chi-square

```
        Shapiro-Wilk normality test

data:  age
W = 0.86268, p-value = 1.853e-12
```

Hypothesis Testing with R Programming

- To perform the *z*-test in R, use the function pnorm()
- For the body temperature example discussed earlier, the *z*-score was −1.
  - For the one-sided hypothesis of the form $H_A: \mu < \mu_0$, we find the lower tail probability of −1 as follows:

```
> pnorm(-1, mean = 0, sd = 1, lower.tail = TRUE)
[1]  0. 1586553
```

- For the BMI example, *z*-score was 1.67.
  - For the one-sided hypothesis of the form $H_A: \mu > \mu_0$, we need to find the upper tail probability of 1.67 as follows:

```
> pnorm(1.67, mean = 0, sd = 1, lower.tail = FALSE)
[1]  0.04745968
```

Hypothesis Testing with R Programming

- When σ^2 is unknown and we need to use the data to estimate it separately, we use the *t*-test to evaluate hypotheses regarding the mean of a normal distribution.
- For the BMI example discussed earlier, the *t*-score was $t = 5.33$.
  - For the one-sided hypothesis of the form $H_A : \mu > \mu_0$, we find the upper tail probability of 5.33 from a *t* distribution with $n-1$ degrees of freedom, where $n = 200$ in this example.
  - We use the pt() function:

```
> pt(5.33, df = 199, lower.tail = FALSE)
[1]  0.0000001324778
```

Statistical Inference for the Relationship Between Two Variables

Relationship Between a Numerical Variable and a Binary Variable

- In general, we can denote the means of the two groups as μ_1 and μ_2.
- The null hypothesis indicates that the population means are equal, $H_0 : \mu_1 = \mu_2$.
- In contrast, the alternative hypothesis is one the following:
  - if $H_A : \mu_1 > \mu_2$, if we believe the mean for group 1 is greater than the mean for group 2.
  - if $H_A : \mu_1 < \mu_2$, if we believe the mean for group 1 is less than the mean for group 2.
  - if $H_A : \mu_1 \neq \mu_2$, if we believe the means are different but we do not specify which one is greater.
- We can also express these hypotheses in terms of the difference in the means:

$$H_A : \mu_1 - \mu_2 > 0,\ H_A : \mu_1 - \mu_2 < 0, \text{ or } H_A : \mu_1 - \mu_2 \neq 0$$

- Then the corresponding null hypothesis is that there is no difference in the population means, $H_0 : \mu_1 - \mu_2 = 0$

Relationship Between a Numerical Variable and a Binary Variable

- Previously, we used the sample mean $\overline{X}$ to perform statistical inference regarding the population mean μ.
- To evaluate our hypothesis regarding the difference between two means, μ_1 - μ_2, it is reasonable to choose the difference between the sample means, $\overline{X}_1$ - $\overline{X}_2$, as our statistic.
- We use μ_{12} to denote the difference between the population means μ_1 and μ_2, and use $\overline{X}_{12}$ to denote the difference between the sample means $\overline{X}_1$ and $\overline{X}_2$:

$$\mu_{12} = \mu_1 - \mu_2 \qquad \overline{X}_{12} = \overline{X}_1 - \overline{X}_2$$

Relationship Between a Numerical Variable and a Binary Variable

- By the Central Limit Theorem,

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \qquad \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

where n_1 and n_2 are the number of observations.

- Therefore,

$$\bar{X}_{12} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

- We can rewrite this as

$$\bar{X}_{12} \sim N(\mu_{12}, SD_{12}^2) \text{ where } SD_{12} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Relationship Between a Numerical Variable and a Binary Variable

- We want to test our hypothesis that $H_A : \mu_{12} \neq 0$ (i.e., the difference between the two means is not zero) against the null hypothesis that $H_0 : \mu_{12} = 0$.
- To use $\overline{X}_{12}$ as a test statistic, we need to find its sampling distribution under the null hypothesis (i.e., its null distribution).
- If the null hypothesis is true, then $\mu_{12} = 0$.
- Therefore, the null distribution of $\overline{X}_{12}$ is

$$\bar{X}_{12} \sim N(0, SD_{12}^2)$$

- As before, however, it is more common to standardize the test statistic by subtracting its mean (under the null) and dividing the result by its standard deviation.

$$Z = \frac{\bar{X}_{12}}{SD_{12}}$$

- where Z is called the z-statistic, and it has the standard normal distribution: $Z \sim N(0, 1)$.

Two-sample z-test

- To test the null hypothesis $H_0 : \mu_{12} = 0$, we determine the z-score,

$$z = \frac{\bar{x}_{12}}{SD_{12}}$$

- Then, depending on the alternative hypothesis, we can calculate the p-value, which is the observed significance level, as:
  - if $H_A : \mu_{12} > 0, \quad p_{\text{obs}} = P(Z \geq z),$
  - if $H_A : \mu_{12} < 0, \quad p_{\text{obs}} = P(Z \leq z),$
  - if $H_A : \mu_{12} \neq 0, \quad p_{\text{obs}} = 2 \times P(Z \geq |z|),$
- The above tail probabilities are obtained from the standard normal distribution.

Example

- Suppose that our sample includes n_1 = 25 women and n_2 = 27 men.
- The sample mean of body temperature is $\overline{x}_1$ = 98.2 for women and $\overline{x}_2$ = 98.4 for men.
- Then, our point estimate for the difference between population means is $\overline{x}_{12}$ = − 0.2.
- We assume that σ_1^2 = 0.8 and σ_2^2 = 1.
- The variance of the sampling distribution is (0.8/25)+(1/27) = 0.07, and the standard deviation is SD_{12} = √0.07 = 0.26.
- The *z*-score is $z = \frac{\bar{x}_{12}}{SD_{12}} = \frac{-0.2}{0.26} = -0.76$

Example

- $H_A : \mu_{12} \neq 0$ and $z = -0.76$.
- Therefore, $p_{\text{obs}} = 2P(Z \geq |-0.76|) = 2 \times 0.22 = 0.44$.
- For the body temperature example, $p_{\text{obs}} = 0.44$ is greater than the commonly used significance levels (e.g., 0.01, 0.05, and 0.1).
- Therefore, the test result is not statistically significant, and we cannot reject the null hypothesis (which states that the population means for the two groups are the same) at these levels.
- That is, any observed difference could be due to chance alone.

Two-Sample t-test

- In practice, SD_{12} is not known since σ_1 and σ_2 are unknown.
- As before, we can use the sample variances S_1^2 and S_2^2 to estimate σ_1^2 and σ_1^2, and take this additional source of uncertainty into account by using t -distributions instead of the standard normal distribution.
- We use s_1^2 and s_2^2 (point estimates for population variances σ_1^2 and σ_1^2) to estimate the standard deviation,

$$SE_{12} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

  where SE_{12} is the standard error of $\overline{X}_{12}$.
- Then, instead of the standard normal distribution, we need to use t-distributions to find *p-values*.
- For this, we can use R or R-Commander.

Two-Sample t-test

- Using the specific value of $\bar{X}_{12}$, which is denoted $\bar{x}_{12}$, as our point estimate for the difference between the two population means, $\mu_{12} = \mu_1 - \mu_2$, along with the standard error SE_{12} of $\bar{X}_{12}$, we find confidence intervals for μ_{12} as follows:

$$[\bar{x}_{12} - t_{crit} \times SE_{12}, \bar{x}_{12} + t_{crit} \times SE_{12}]$$

  where t_{crit} is the t-critical value from a t-distribution for the desired confidence level c.

- When comparing the population means for two groups, the formula for finding the degrees of freedom is as follows:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Two-Sample t-test

- For testing a hypothesis regarding $\mu_{12} = \mu_1 - \mu_2$ when the population variances are unknown, we follow similar steps as above, but we use SE_{12} instead of SD_{12} and use the following *t* -statistic instead of the *z*-statistic to account for the additional source of uncertainty involved in estimating the population variances:

$$T = \frac{\bar{X}_{12}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

where $\overline{X}_{12} = \overline{X}_1 - \overline{X}_2$ as before.

Two-Sample t-test

- Using the observed data, we obtain $\bar{x}_{12} = \bar{x}_1 - \bar{x}_2$ as the observed value of $\bar{X}_{12}$. We also use the observed data to obtain s_1 and s_2 as the observed values of sample variances.
- Then, we calculate the observed value of the test statistic T as follows:

$$t = \frac{\bar{x}_{12}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\bar{x}_{12}}{SE_{12}}$$

  which is called the t -score.
- Depending on the alternative hypothesis, we calculate p_{obs} as
  - if $H_A : \mu_{12} > 0$, $p_{\text{obs}} = P(T \geq t)$,
  - if $H_A : \mu_{12} < 0$, $p_{\text{obs}} = P(T \leq t)$,
  - if $H_A : \mu_{12} \neq 0$, $p_{\text{obs}} = 2 \times P(T \geq |t|)$,

  where T has a t -distribution with the degrees of freedom obtained as above

Example

- For the body temperature example, suppose that the sample variances based on our sample of n_1 = 25 women and n_2 = 27 men are s_1^2= 1.1 and s_2^2= 1.2, respectively.
- The standard error of $\overline{X}_{12}$ is

$$SE_{12} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{1.1}{25} + \frac{1.2}{27}} = 0.3$$

- Degrees of freedom is

Example

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{1.1}{25} + \frac{1.2}{27}\right)^2}{\frac{1}{25 - 1}\left(\frac{1.1}{25}\right)^2 + \frac{1}{27 - 1}\left(\frac{1.2}{27}\right)^2} = 49.9$$

- To find the corresponding t_{crit}, we follow similar steps as before.
- Suppose that we are interested in 95% confidence interval for μ_{12}.
- We find t_{crit} from the t –distribution with $df = 49.9$ degrees of freedom.
- In R-Commander,
  - click *Distributions* → *t distribution* → *t quantiles*.
  - Then enter (1 − 0.95)/2 = 0.025 for Probabilities, 49.9 for Degrees of freedom, and check the option Upper tail.
- The corresponding t-critical value is 2.01.

Example

- This results in the following 95% confidence interval:

$$[\bar{x}_{12} - t_{crit}\times SE_{12}, \bar{x}_{12} + t_{crit}\times SE_{12}]$$

$$[-0.2 - 2.01\times 0.30, -0.2 + 2.01\times 0.30] = [-0.80, 0.40].$$

- Therefore, at 0.95 confidence level, we believe that the true difference between the two means falls between −0.80 and 0.40.
- The t –score is

$$t = \frac{\bar{x}_{12}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\bar{x}_{12}}{SE_{12}} = \frac{-0.2}{0.3} = -0.67$$

Example

- The alternative hypothesis is $H_A : \mu_{12} \neq 0$.
- Using the t-distribution with $df = 49.9$ degrees of freedom, the upper tail probability of $|-0.67| = 0.67$ is $P(T > 0.67) = 0.25$.
- The observed significance level is $p_{obs} = 2 \times 0.25 = 0.50$, which is considered to be large (compared to commonly used significance levels).
- Therefore, the result is not statistically significant, and we cannot reject the null hypothesis, which indicates that the two populations (men and women) have the same mean body temperature.

Analysis of Variance (ANOVA)

ANOVA

- The process of evaluating hypotheses regarding the group means of multiple populations is called the Analysis of Variance (ANOVA).
- ANOVA models generalize the *t*-test and are used to compare the means of multiple groups identified by a categorical variable with more than two possible categories.
- Since we are only considering one factor only, this method is specifically called one- way ANOVA.
- An ANOVA with two factors is called a two-way ANOVA.
- In general, the between-groups variation is denoted as SS_B and calculated by

$$SS_B = \sum_{i=1}^{k} n_i \, (\overline{y}_i - \overline{y})^2$$

where k is the number of groups

ANOVA

- The within-groups variation is denoted as SS_W and calculated by

$$SS_w = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2$$

- We measure the total variation in Y by

$$SS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y})^2$$

- The total variation SS is equal to the sum of the between-groups variation SS_B and the within-groups variation SS_W,

$$SS = SS_B + SS_W.$$

- The total variation can be attributed partly to the variation within groups and partly to the variation between groups. SS_B is interpreted as the part of total variation SS that is associated with (and can be explained by) the factor variable X (e.g., syndrome type).
- In contrast, SS_W is regarded as the unexplained part of total variation and is regarded as random.

ANOVA

- Let us denote the overall population mean of Y as μ and group-specific population means as μ_1 , . . . , μ_4 .
- Then we can express the null hypothesis of no difference in means between the groups as

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$$

- The alternative hypothesis H_A is that at least one of the group means μ_i is different from the mean μ.
- The test statistic for examining the null hypothesis is called F-statistic (more specifically, ANOVA F -statistic) and is defined as

$$F = \frac{SS_B/(k-1)}{SS_W/(n-k)}$$

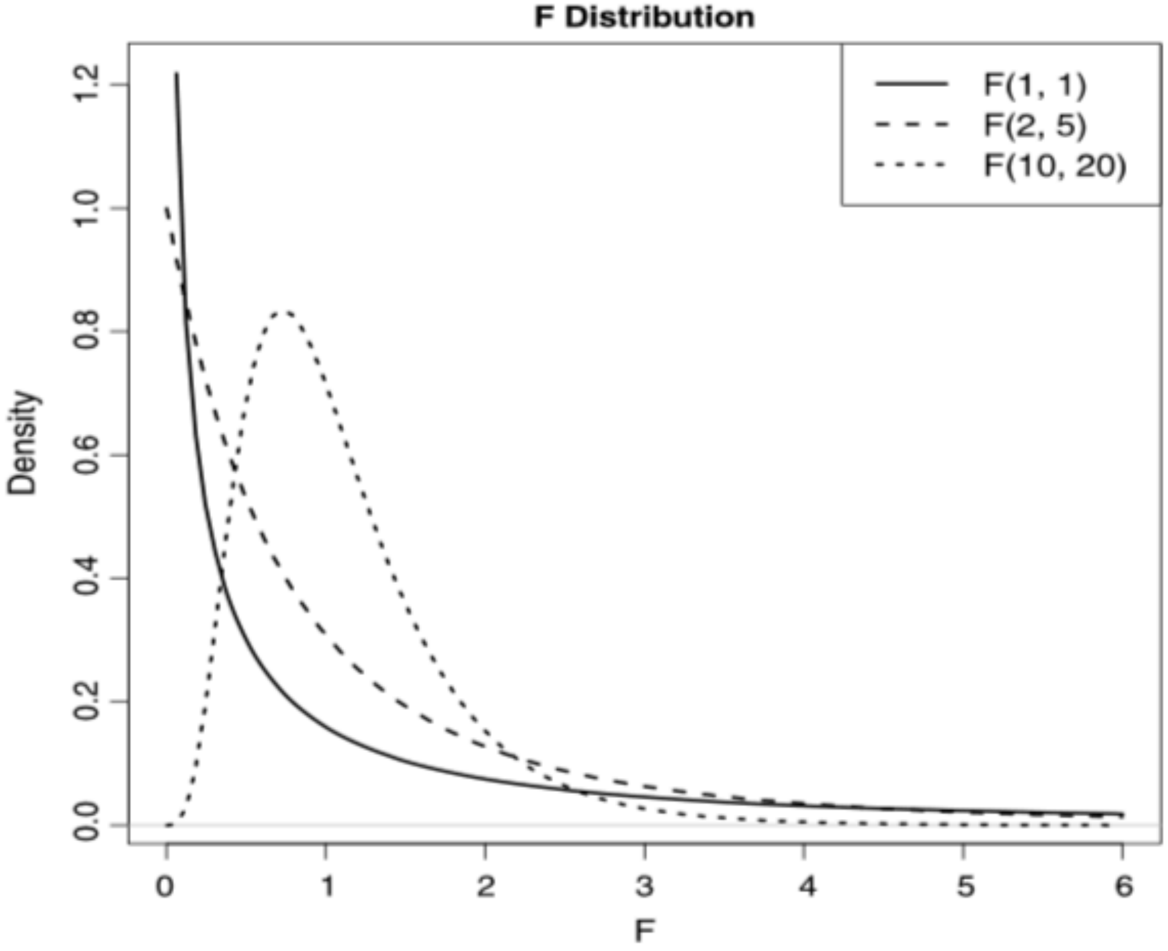
where n is the total sample size, and k is the number of groups.
- The numerator is called the mean square for groups, and the denominator is called the mean square error (MSE).

ANOVA

- For the one-way ANOVA, the F-statistic has: $F(df_1 = k - 1, df_2 = n - k)$ distribution under the null hypothesis (i.e., assuming that the null hypothesis is true).
- The F-distribution, which is a continuous probability distribution, is very important for hypothesis testing.
- It is specified by two parameters, df_1 and df_2, and is denoted as F(df_1, df_2).
- We refer to df_1 and df_2 as the numerator degrees of freedom and denominator degrees of freedom, respectively.
- Both parameters must be positive.

ANOVA

- The following figure shows the pdf of F-distribution for different values of df_1 and df_2.



Example

- As an example, we analyze the Cushings data set, which is available from the MASS package.
  - Cushing's syndrome is a hormone disorder associated with high level of cortisol secreted by the adrenal gland.
- The *Type* variable in the data set shows the underlying type of syndrome, which can be one of four categories:
  - adenoma (a), bilateral hyperplasia (b), carcinoma (c), and unknown (u).

Example

- Our objective is to find whether the four groups are different with respect to urinary excretion rate of Tetrahydrocortisone.
- We denote by Y the urinary excretion rate of Tetrahydrocortisone and by X the *Type* variable,
  - where $X = 1$ for Type=a, $X = 2$ for Type=b, $X = 3$ for Type=c, and $X = 4$ for Type=u.
- Then, our objective could be defined as investigating whether the *mean* of the response variable Y differs for different values (levels) of the factor X.

Example

- Denote the individual observations as y_{ij} : the urinary excretion rate of Tetrahydrocortisone of the j th individual in group i.
- Total number of observations is $n = 27$,
- The number of observations in each group is

  $n_1 = 6$, $n_2 = 10$, $n_3 = 5$, and $n_4 = 6$.
- The overall (for all groups) observed sample mean for the response variable is $\overline{y} = 10.46$.
- We also find the group specific means, by clicking (in R-Commander) *Statistics→Summaries→Numerical summaries*

  $\bar{y}_1 = 3.0$, $\bar{y}_2 = 8.2$, $\bar{y}_3 = 19.7$, and $\bar{y}_4 = 14.0$.
- The degrees of freedom parameters are

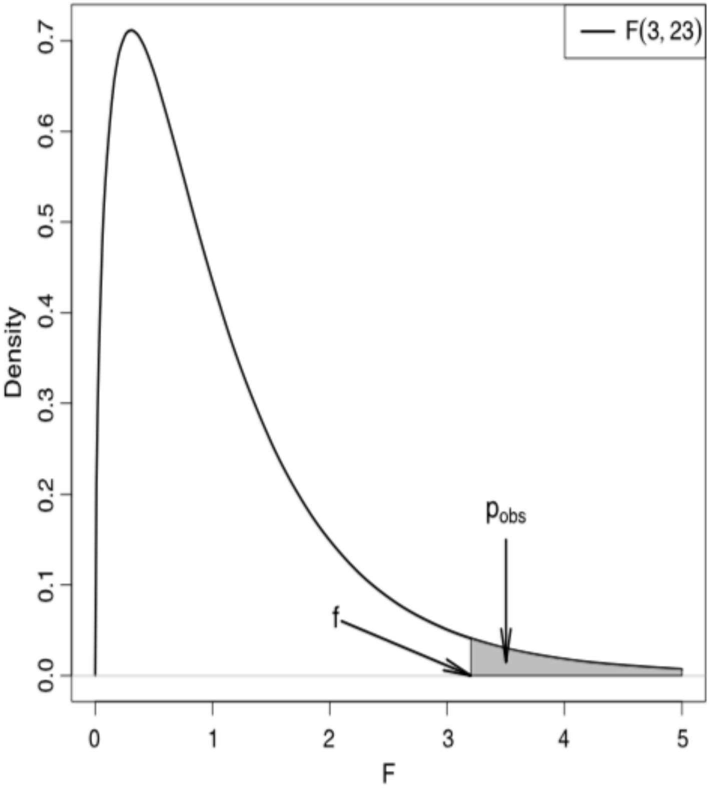
  $df_1 = 4-1 = 3$ and $df_2 = 27 - 4 = 23$.

Example

- $SS_B = 893.5$ and $SS_W = 2123.6$.
- The observed value of F-statistic is $f = 3.2$ given under the column labeled F value.
- The resulting *p*-value is then 0.04.
- Therefore, we can reject H_0 at 0.05 significance level (but not at 0.01) and conclude that the differences among group means for urinary excretion rate of Tetrahydrocortisone are statistically significant (at 0.05 level).

Example

- For plotting the *F*(3, 23) distribution using R-Commander, click *Distribution → Continuous distributions→ F distribution Plot F distribution*.
- In R commad line: pf(3.2, df1=3, df2=23, lower.tail=F)
- Set the *Numerator degrees of freedom* to 3 and the *Denominator degrees of freedom* to 23.



- The density plot of *F*(3, 23)-distribution.
- This is the distribution of *F*-statistic for the Cushings data assuming that the null hypothesis is true.
- The observed value of the test statistic is *f* = 3.2, and the corresponding *p*-value is shown as the *shaded area* above 3.2

ANALYSIS OF CATEGORICAL VARIABLES

ANALYSIS OF CATEGORICAL VARIABLES

- Pearson's χ^2 (chi-squared) test is used to test hypotheses regarding the distribution of a categorical variable or the relationship between two categorical variables.
- Pearson's χ^2 test uses a test statistic, which we denote as Q, to measure the discrepancy between the observed data and what we expect to observe under the null hypothesis.
- Higher levels of discrepancy between data and H_0 results in higher values of Q.
- We use q to denote the observed value of Q based on a specific sample of observed data.

Pearson's χ^2 Test for One Categorical Variable

- Let us denote the binary variable of interest as X, based on which we can divide the population into two groups depending on whether $X = 1$ or $X = 0$.
- Further, suppose that the null hypothesis H_0 states that the probability of group 1 is $\boldsymbol{\mu_{01}}$ and the probability of group 2 is $\boldsymbol{\mu_{02}}$.
  - Here $\boldsymbol{\mu_{02} = 1 - \mu_{01}}$.
- If the null hypothesis is true, we expect that, out of n randomly selected individuals, $\boldsymbol{E_1 = n\mu_{01}}$ belong to the first group, and $\boldsymbol{E_2 = n(1 - \mu_{01})}$ belong to the second group.
- We refer to E_1 and E_2 as the expected frequencies under the null.
- We refer to the observed number of people in each group as the observed frequencies and denote them $\boldsymbol{O_1}$ and $\boldsymbol{O_2}$ for group 1 and group 2, respectively.

Pearson's χ^2 Test for One Categorical Variable

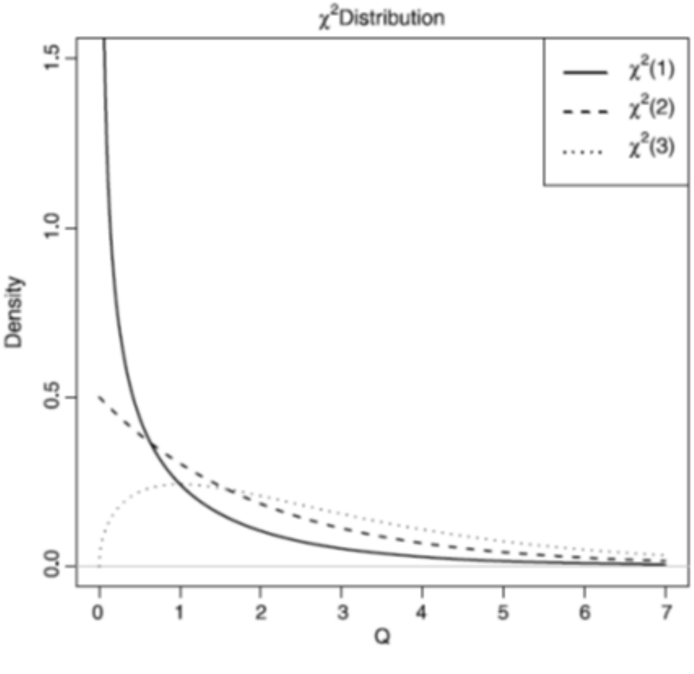
- Pearson's $\chi 2$ test measures the discrepancy between the observed data and the null hypothesis based on the difference between the observed and expected frequencies as follows:

$$Q = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

- The value of Q will be zero only when the observed data matches our expectation under the null exactly.
- When there is some discrepancy between the data and the null hypothesis, Q becomes greater than zero.
- The higher discrepancy between our data and what is expected under H_0, the larger Q and therefore the stronger the evidence against H_0.

Pearson's χ^2 Test for One Categorical Variable

- If the null hypothesis is true, then the approximate distribution of Q is χ^2.
- Like the t-distribution, the χ^2-distribution is commonly used for hypothesis testing and denoted $\chi^2(df)$.
- The plot of the pdf for a χ^2 distribution with various degrees of freedom



- The observed significance level p_{obs} is calculated using the χ^2 distribution with 1 degree of freedom.
- This corresponds to the upper tail probability of q from the $\chi^2(1)$ distribution.

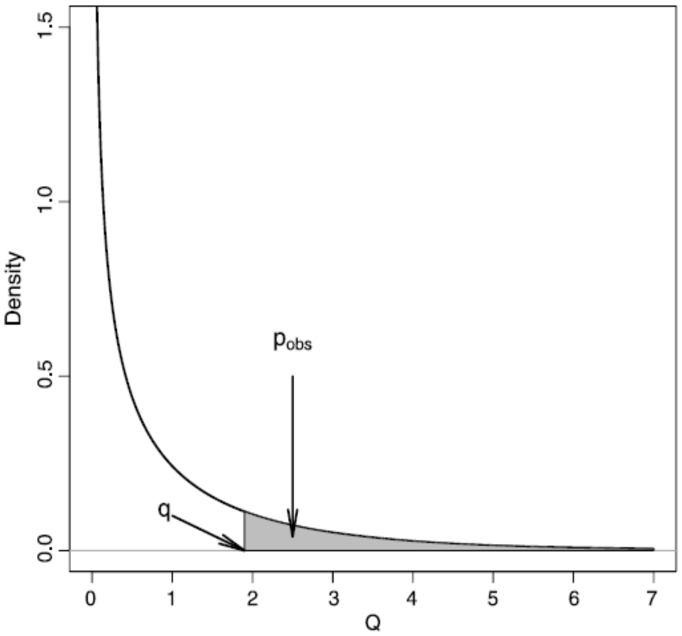
Example

- We use the heart attack survival rate (i.e., the probability of survival after heart attack) within one year after hospitalization.
- Suppose that $\boldsymbol{H_0}$ specifies that the probability of surviving is $\mu_{01} = 0.70$ and the probability of not surviving is $\mu_{02} = 0.30$.
- If we take a random sample of size $n = 40$ from the population, we expect that $E_1 = 0.70 \times 40 = 28$ and $E_2 = 0.30 \times 40 = 12$.
- Now suppose that the observed number of people in each group as the observed frequencies: $O_1 = 24$ and $O_2 = 16$.
- For the heart attack survival example, the observed value of the test statistic is

$$q = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} = \frac{(24 - 28)^2}{28} + \frac{(16 - 12)^2}{12} = 1.90$$

- The $p_{obs} = P(Q \geq 1.90) = 0.17$ is obtained from a χ^2 distribution with 1 degree of freedom

Example



- The sampling distribution for Q under the null hypothesis: $Q \sim \chi^2(1)$.
- The p-value is the upper tail probability of observing values as extreme or more extreme than $q = 1.90$

- Therefore, the results are not statistically significant, and we cannot reject the null hypothesis at commonly used significance levels (e.g., 0.01, 0.05, and 0.1).
- In this case, we believe that the difference between observed and expected frequencies could be due to chance alone.

Categorical Variables with Multiple Categories

- Pearson's χ^2 test can be generalized to situations where the categorical random variable can take more than two values.
- In general, for a categorical random variable with k possible categories, we calculate the test statistic Q as

$$Q = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \, ,$$

- The approximate distribution of Q is $\chi 2$ with the degrees of freedom equal to $df = k - 1$.
- Therefore, to find p_{obs}, we calculate the upper tail probability of q (the observed value of Q) from the $\chi^2(k - 1)$ distribution.

Example

- Suppose that we monitor heart attack patients for one year and divide them into three groups:
  - patients who did not have another heart attack and survived,
  - patients who had another heart attack and survived,
  - patients who did not survive.
- Suppose that $\mu_{01} = 0.5$, $\mu_{02} = 0.2$, and $\mu_{03} = 0.3$.
- The expected frequencies of each category for a sample of $n = 40$:

$$E_1 = 0.5\times 40 = 20,\ E_2 = 0.2\times 40 = 8,\ E_3 = 0.3\times 40 = 12.$$

- This time, suppose that the actual observed frequencies based on a sample of size $n = 40$ for the three groups are

$$O_1 = 13,\ O_2 = 11,\ O_3 = 16.$$

Example

- The amount of discrepancy:

$$Q = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3}$$

- The observed value of this test statistic:

$$q = \frac{(13 - 20)^2}{20} + \frac{(11 - 8)^2}{8} + \frac{(16 - 12)^2}{12} = 4.91$$

- Using R-Commander, we find $p_{\text{obs}} = P(Q \geq 4.91) = 0.086$ using the χ^2 distribution with 2 degrees of freedom.
- Therefore, we can reject the null hypothesis at 0.1 level but not at 0.05 level. Note: In R: *pchisq(4.91, df=2, lower.tail=F)* to get the p_{obs}
- At the 0.1 significance level, we can conclude that the difference between observed and expected frequencies is statistically significant, and it is probably not due to chance alone.

Pearson's χ^2 Test of Independence

- We now discuss the application of Pearson's χ^2 test for evaluating a hypothesis regarding possible relationship between two categorical variables.
- More specifically, we measure the difference between the observed frequencies and expected frequencies under the null.
- The null hypothesis in this case states that the two categorical random variables are independent.
  - For two independent random variables, the joint probability is equal to the product of their individual probabilities.
    - In what follows, we use this rule to find the expected frequencies.

Pearson's χ^2 Test of Independence

- We use the following general form of Pearson's χ^2 test, which summarizes the differences between the expected frequencies (under the null hypothesis) and the observed frequencies over all cells of the contingency table:

$$Q = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} and E_{ij} are the observed and expected values in the ith row and j th column of the contingency table.

- The double sum simply means that we add the individual measures of discrepancies for cells by going through all cells in the contingency table.

Pearson's χ^2 Test of Independence

- As before, higher values of Q provide stronger evidence against H_0.
- For $I \times J$ contingency tables (i.e., I rows and J columns), the Q statistic has approximately the χ^2 distribution with $(I-1)\times(J-1)$ degrees of freedom under the null.
- Therefore, we can calculate the observed significance level by finding the upper tail probability of the observed value for Q, which we denote as q, based on the $\chi2$ distribution with $(I-1)\times(J-1)$ degrees of freedom.

Example 1

- The probability that the mother is smoker (i.e., smoke =1) and the baby has low birthweight (i.e., low =1) is the product of smoker and low-birthweight probabilities.
- For the baby weight example, we can summarize the observed and expected frequencies in the contingency tables.

	Observed frequency				Expected frequency	
	Normal	Low			Normal	Low
Nonsmoking	86	29		Nonsmoking	79.1	35.9
Smoking	44	30		Smoking	50.9	23.1

Example 1

- Then Pearson's test statistic is

$$Q = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(O_{22} - E_{22})^2}{E_{22}}$$

$$q = \frac{(86 - 79.1)^2}{79.1} + \frac{(29 - 35.9)^2}{35.9} + \frac{(44 - 50.9)^2}{50.9} + \frac{(30 - 23.1)^2}{23.1} = 4.9$$

- Since the table has $I = 2$ rows and $J = 2$ columns, the approximate null distribution of Q is χ^2 with $(2-1)\times(2-1) = 1$ degrees of freedom.
- Consequently, the observed p-value is the upper tail probability of 4.9 using the $\chi^2(1)$ distribution.
- We find $p_{\text{obs}} = P(Q \geq 4.9) = 0.026$.
- Therefore, at the 0.05 significance level (but not at 0.01 level), we can reject the null hypothesis that the mother's smoking status and the baby's birthweight status are independent.

Example 2

- Suppose that we would like to investigate whether the race of mothers is related to the risk of having babies with low birthweight.
- The race variable can take three values: 1 for white, 2 for African-American, and 3 for others.
- As before, the low variable can take 2 possible values: 1 for babies with birthweight less than 2.5 kg and 0 for other babies.
- Therefore, all possible combinations of race and low can be presented by a 3 × 2 contingency table.
- The following Table provides the observed frequency of each cell and the expected frequency of each cell if the null hypothesis is true.

Example 2

	Observed frequency				Expected frequency	
Groups	Normal (low=0)	Low (low=1)		Groups	Normal (low=0)	Low (low=1)
1	73	23		1	66	30
2	15	11		2	18	8
3	42	23		3	46	21

- For example, there are 73 babies in the first row and first column.
- This is the number of babies in the intersection of race = 1 (mother is white) and low = 0 (having a baby with normal birthweight).
- If the null hypothesis is true, the expected number of babies in this cell would have been 66.

Example 2

- The observed value of the test statistic Q is obtaned as $q = 5.0$ using the following equations.
- The distribution of Q under the null hypothesis is χ^2 with

  $(3 - 1) \times (2 - 1) = 2$ degrees of freedom.
- To find the corresponding p-value, we need to find the probability of observing values as or more extreme than 5.0.
- This is the upper-tail probability of 5 from the $\chi^2(2)$ distribution: $p_{\text{obs}} = P(Q \geq 5)$.

$$Q = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(O_{22} - E_{22})^2}{E_{22}} + \frac{(O_{31} - E_{31})^2}{E_{31}} + \frac{(O_{32} - E_{32})^2}{E_{32}}$$

- The value of p_{obs} is 0.08.
- We can reject the null hypothesis at 0.1 level but not at 0.05 level.
- At 0.05 level, the relationship between the two variables (i.e., race of mothers and birthweight status) is not statistically significant.

Regression Analysis

Regression Analysis

- The modeling of the relationship between a response variable and a set of explanatory variables is one of the most widely used of all statistical techniques.
  - We refer to this type of modeling as regression analysis.
- A regression model provides the user with a functional relationship between the response variable (often called as "dependent variable") and explanatory variables (often called as "independent variables") that allows the user to determine which of the explanatory variables have an effect on the response.
  - The regression model allows the user to explore what happens to the response variable for specified changes in the explanatory variables.
    - {For example, financial officers must predict future cash flows based on specified values of interest rates, raw material costs, salary increases, and so on}

Regression Analysis

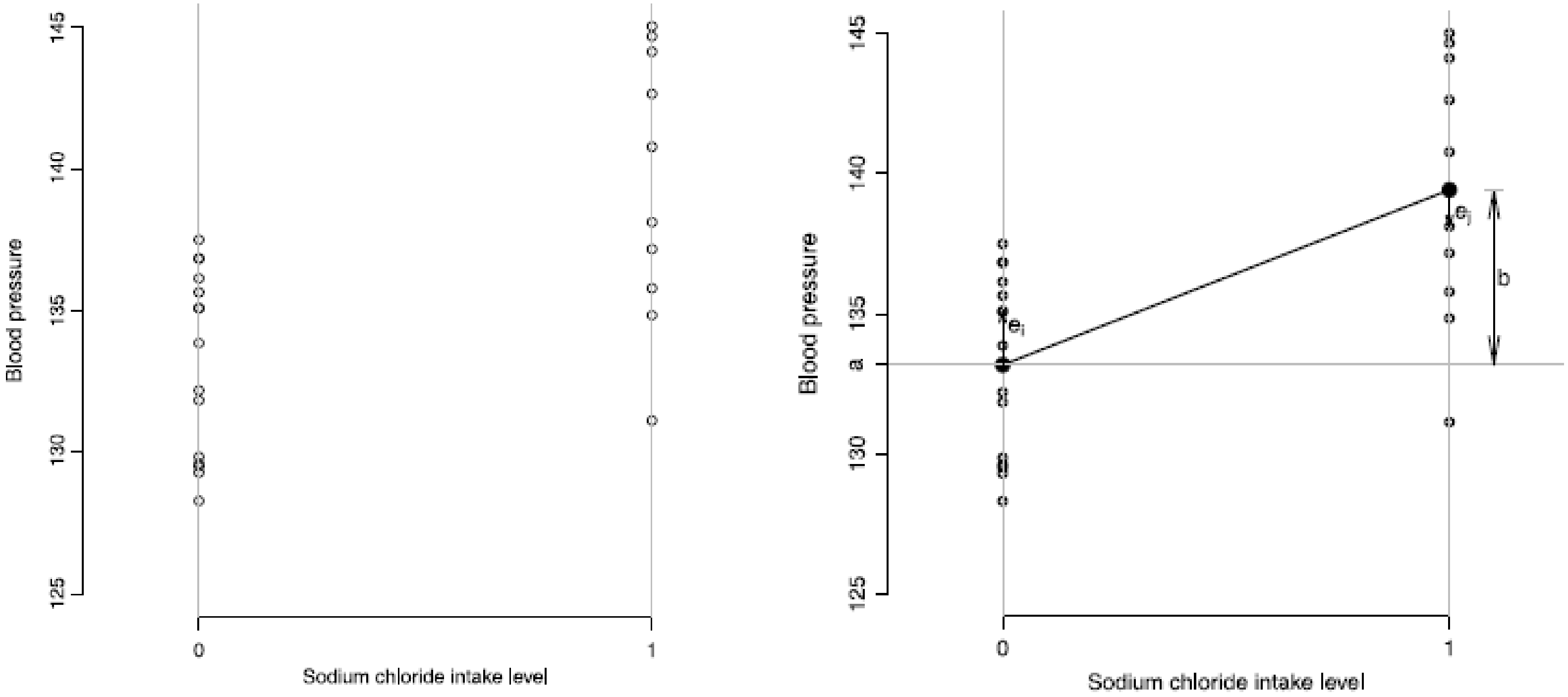
- We now discuss linear regression models for either testing a hypothesis regarding the relationship between one or more explanatory variables and a response variable, or predicting unknown values of the response variable using one or more predictors
  - We use X to denote explanatory variables and Y to denote response variables.
- We start by focusing on problems where the explanatory variable is binary.
  - As before, the binary variable X can be either 0 or 1.
- We then continue our discussion for situations where the explanatory variable is numerical.

Linear Regression Models with One Binary Explanatory Variable

- Suppose that we want to investigate the relationship between sodium chloride (salt) consumption (low vs. high consumption) and blood pressure among elderly people (e.g., above 65 years old).
- The next figure shows the dot plot along with sample means, shown as black circles, for each group.
- We connect the two sample means to show the overall pattern for how blood pressure changes from one group to another.

Linear Regression Models with One Binary Explanatory Variable

- The dot plot for systolic blood pressure for 25 elderly people (left panel), where 15 people follow a low sodium chloride diet ($X = 0$), and 10 people follow a high sodium chloride diet ($X = 1$)
- The dot plot for systolic blood pressure for 25 elderly people (right panel).
  - Here, the sample means among the low and high sodium chloride diet groups are shown as black circles. A straight line connects the sample means. The line intercepts the vertical axis at a = 133.17 and has slope b = 6.25



Linear Regression Models with One Binary Explanatory Variable

- Using the intercept ***a*** and slope ***b***, we can write the equation for the straight line that connects the estimates of the response variable for different values of *X* as follows:

$$\hat{y} = a + bx$$

  - The constant (intercept) term *a* is interpreted as the predicted value of *y* when $x = 0$.
  - The slope *b* of the line is the predicted change in *y* when there is a one-unit change in *x*.

- The slope is also known as the regression coefficient of *X*.
  - For the given example,

$$\hat{y} = 133.17 + 6.25x$$

    - We expect that on average the blood pressure increases by 6.25 units for one unit increase in *X*.
    - In this case, one unit increase in *X* from 0 to 1 means moving from low to high sodium chloride diet group.

Linear Regression Models with One Binary Explanatory Variable

- For an individual with $x = 0$, the estimate according to the above regression line is

$$\hat{y} = a + b \times 0 = a = \hat{y}_{x=0}$$

which is the sample mean for the first group.

- For an individual with $x = 1$, the estimate according to the above regression line is

$$\hat{y} = a + b \times 1 = a + b = \hat{y}_{x=0} + \hat{y}_{x=1} - \hat{y}_{x=0} = \hat{y}_{x=1}$$

- We refer to the difference between the observed and estimated values of the response variable as the residual.
  - For individual *i*, we denote the residual e_i and calculate it as follows:

$$e_i = y_i - \hat{y}_i$$

Linear Regression Models with One Binary Explanatory Variable

- For instance, if someone belongs to the first group, her estimated blood pressure is

$$\hat{y}_i = \text{a} = 133.17$$

- Now if the observed value of her blood pressure is y_i = 135.08, then the residual is

$$e_i = 135.08 – 133.17 = 1.91$$

- By rearranging the terms in the equation $e_i = y_i - \hat{y}_i$, we can write the observed value y_i in terms of the estimate obtained from the regression line and the corresponding residual,

$$y_i = \hat{y}_i + e_i$$

- For individual i, whose values of the explanatory variable and the response variable are x_i and y_i, respectively, the estimated value of the response variable, denoted as $\hat{y}_i$, is

$$\hat{y}_i = a + bx_i$$

- So, the observed value y_i can be given as

$$y_i = a + bx_i + e_i$$

The linear relationship

- The linear relationship between *Y* and *X* in the entire population can be presented in a similar form,

$$Y = \alpha + \beta X + \varepsilon$$

- where α is the intercept, and β is the slope of the regression line , ε is called the **error term**, representing the difference between the estimated and the actual values of *Y* in the population.
- We refer to the above equation as the **linear regression model**.
  - We refer to α and β as the **regression parameters**.
  - More specifically, β is called the **regression coefficient** for the explanatory variable.
  - The process of finding the regression parameters is called **fitting** a regression model to the data.

Statistical Inference Using Simple Linear Regression Models

- Using the regression line, we can estimate the unknown value of the response variable for members of the population who did not participate in our study.
- In this case, we refer to our estimates as **predictions**.
  - For example, we can use the linear regression model we built in the previous example to predict the value of blood pressure for a person with high sodium chloride diet (i.e., $x = 1$),

$$
\begin{aligned}
\hat{y} &= 133.17 + 6.25 \times x \\
&= 133.17 + 6.25 \times 1 \\
&= 139.42
\end{aligned}
$$

Residual sum of squares

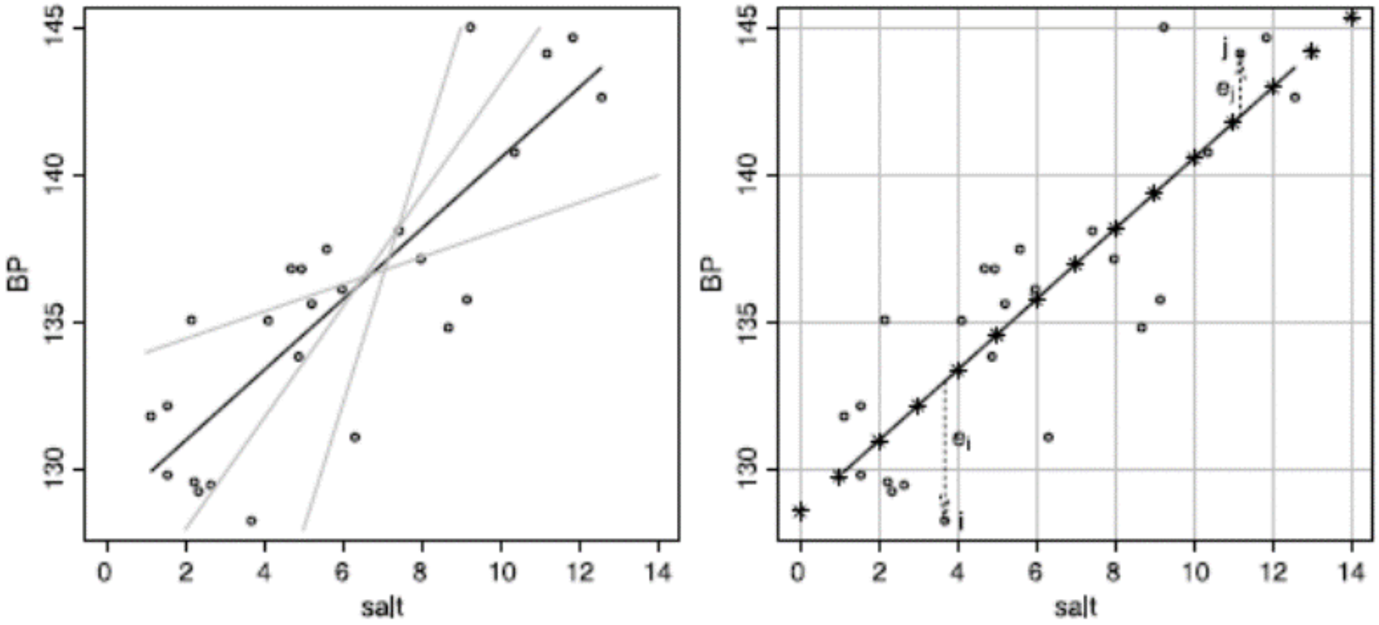
- As a measure of discrepancy between the observed values and those estimated by the line, we calculate the Residual Sum of Squares (*RSS*):

$$RSS = \sum_{i}^{n} e_i^2$$

- Here, e_i is the residual of the *i*th observation, and *n* is the sample size.
- The square of each residual is used so that its sign becomes irrelevant.

One Numerical Explanatory Variable

- We now discuss simple linear regression models (i.e., linear regression with only one explanatory variable), where the explanatory variable is numerical.



- *Left panel*: Scatterplot of blood pressure by daily sodium chloride intake along with some candidate lines for capturing the overall relationship between the two variables.
  - The *black line* is the least-squares regression line.
- *Right panel*: The least-squares regression line for the relationship between blood pressure and sodium chloride intake.
  - The *vertical arrows* show the residuals for two observations.
  - The *stars* are the estimated blood pressure for daily sodium chloride intakes from 0 to 14 grams

One Numerical Explanatory Variable

- Among all possible lines we can pass through the data, we choose the one with the smallest sum of squared residuals.
  - The resulting line is called **the least-squares regression line**.
- First, we find the slope of regression line using the sample correlation coefficient r between the response variable Y and the explanatory variable X,

$$b = r\frac{s_y}{s_x}$$

  - Here, s_y is the sample standard deviation of Y , and s_x is the sample standard deviation of X.
    - Note that since s_x and s_y are always positive, the sign of b is the same as the sign of the correlation coefficient: $b > 0$ for positively correlated random variables, and $b < 0$ for negatively correlated variables.

One Numerical Explanatory Variable

- When $r = 0$ (i.e., the two variables are not linearly related), then $b = 0$.
- After finding the slope, we find the intercept as follows:

$$a = \overline{y} - b\overline{x}$$

where $\overline{y}$ and $\overline{x}$ are the sample means for Y and X, respectively.

- Then the least-squares regression line with intercept a and slope b can be expressed as

$$\hat{y} = a + bx$$

Example

- For the blood pressure example,
  - the sample correlation coefficient is $r = 0.84$;
  - the sample standard deviation of blood pressure is $s_y = 4.94$,
  - the sample standard deviation of sodium chloride intake is $s_x = 3.46$.
- Therefore,

$$b = 0.84 \times 4.94/3.46 = 1.20.$$

- For the observed data,
  - the sample means are $\overline{y} = 135.68$ and $\overline{x} = 5.90$.
- Therefore,

$$a = 135.68 - 1.20 \times 5.90 = 128.60.$$

- The linear regression model can be written as

$$\hat{y} = 128.60 + 1.20x.$$

Example

- We can now use this model to estimate the value of the response variable.
- For the individual *i* in the right panel of the figure in slide 16,
  - the amount of daily sodium chloride intake is $x_i = 3.68$.
- The estimated value of the blood pressure for this person is

$$\hat{y}_i = 128.60+1.20\times3.68 = 133.02.$$

- The actual blood pressure for this individual is

$$y_i = 128.3$$

- The residual therefore is

$$e_i = y_i - \hat{y}_i = 128.3-133.02=-4.72$$

Example

- We can also use our model for predicting the unknown values of the response variable (i.e., blood pressure) for all individuals in the target population.
  - For example, if we know the amount of daily sodium chloride intake is $x = 7.81$ for an individual, we can predict her blood pressure as follows:

$$\hat{y} = 128.60 + 1.20 \times 7.81 = 137.97$$

- Of course, the actual value of the blood pressure for this individual would be different from the predicted value.
  - The difference between the actual and predicted values of the response variable is called the model **error** and is denoted as ε.
    - In fact, the residuals are the observed values of ε for the individuals in our sample.

Estimating model parameters

- As an alternative way, the least-squares estimates of slope and intercept can be obtained as follows:

  $$\beta = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \alpha = \overline{y} - \beta\overline{x}$$

  where

  $S_{xy} = \sum_i (x_i - \overline{x})(y_i - \overline{y})$ and $S_{xx} = \sum_i (x_i - \overline{x})^2$

- Thus, S_{xy} is the sum of x deviations times y deviations and S_{xx} is the sum of x deviations squared.

Example

- In the road resurfacing example

– Cost y_i (in thousands of dollars):	6.0	14.0	10.0	14.0	26.0
– Mileage x_i (in miles):	1.0	3.0	4.0	5.0	7.0

- For the road resurfacing data, *n* =5 and $\sum x_i = 1.0 + 3.0 + 4.0 + 5.0 + 7.0 = 20.0$
- So $\bar{x} = \frac{20.0}{5} = 4.0.$
- Similarly $\sum y_i = 70.0, \bar{y} = \frac{70.0}{5} = 14.0$
- Also,

$$S_{xx} = \sum_i (x_i - \bar{x})^2 = (1.0 - 4.0)^2 + \cdots \ldots + (7.0 - 4.0)^2 = 20.00$$

Example

- And

$$S_{xy} = \sum_i (x_i - \overline{x})(y_i - \overline{y})$$
$$= (1.0 - 4.0)(6.0 - 14.0) + \cdots \ldots (7.0 - 4.0)(26.0 - 14.0)$$
$$= 60.0$$

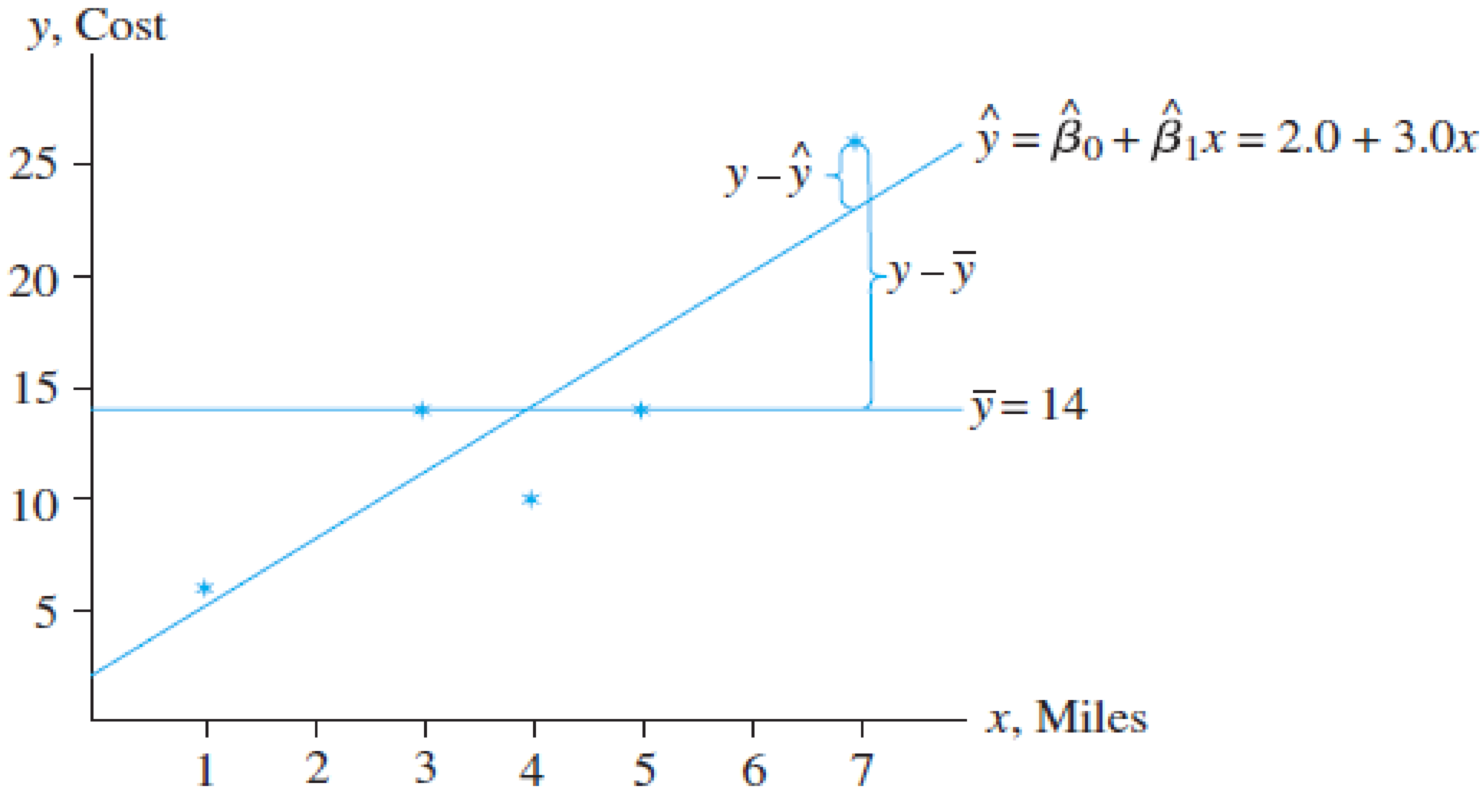
- Thus

$$\beta = \frac{60.0}{20.0} = 3.0$$
$$\alpha = 14.0 - (3.0)(4.0) = 2.0$$

- From the value $= 3.0$, we can conclude that the estimated average increase in cost for each additional mile is $3,000.

Example

- Deviations from the least-squares line from the mean



Statistical inference using regression models

- We can use R or R-Commander to find the least-squares regression line.
- The slope of the regression line plays an important role in evaluating the relationship between the response variable and explanatory variable(s).
- We can also use this regression line to predict the unknown value of the response variable.

Confidence Interval for Regression Coefficients

- We can find the confidence interval for the population regression coefficient as follows:

$$[b - t_{\mathrm{crit}} \times SE_b, b + t_{\mathrm{crit}} \times SE_b].$$

- For simple (i.e., one predictor) linear regression models, SE_b is obtained as follows:

$$SE_b = \frac{\sqrt{RSS/(n-2)}}{\sqrt{\sum_i (x_i - \bar{x})^2}}.$$

- The corresponding t_{crit} is obtained from the *t*-distribution with n - 2 degrees of freedom.

Confidence Interval for Regression Coefficients

- For the blood pressure example,
  - the sample size is $n = 25$.
- Therefore, we use the t -distribution with $25 - 2 = 23$ degrees of freedom.
- If we set the confidence level to 0.95,
  - then $t_{\text{crit}} = 2.07$,
    - which is obtained from the t -distribution with 23 degrees of freedom by setting the upper tail probability to $(1-0.95)/2 = 0.025$.
- Therefore,
  - the 95% confidence interval for β is

$$[6.25-2.07\times1.59,\ 6.25+2.07\times1.59] = [2.96,\ 9.55]$$

Hypothesis testing

- To assess the null hypothesis that the population regression coefficient is zero, which is interpreted as no linear relationship between the response variable and the explanatory variable, we first calculate the *t*-score.

$$t = \frac{b}{SE_b}$$

- Then, we find the corresponding *p*-value as follows:
  - if $H_A : \beta < 0,$ $p_{\text{obs}} = P(T \leq t),$
  - if $H_A : \beta > 0,$ $p_{\text{obs}} = P(T \geq t),$
  - if $H_A : \beta \neq 0,$ $p_{\text{obs}} = 2 \times P(T \geq |t|),$

  where *T* has the *t*-distribution with *n* - 2 degrees of freedom

Hypothesis testing

- In the blood pressure example,
  - the estimate of the regression coefficient was $b = 6.25$,
  - the standard error was $SE_b = 1.59$.
- Therefore,

$$t = b / SE_b = 6.25 / 1.59 = 3.93.$$

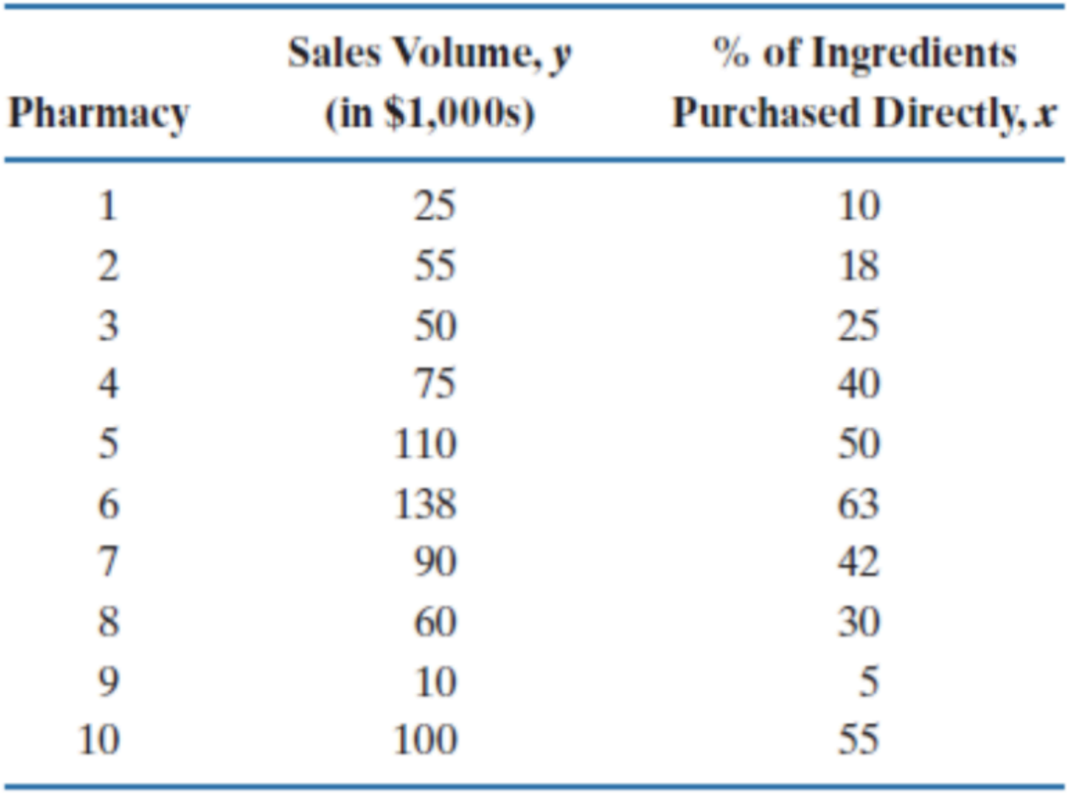
- If $H_A : \beta \neq 0$ (which is the common form of the alternative hypothesis),
  - we find the p-value by calculating the upper tail probability of $|3.93| = 3.93$ from the t -distribution with $25 - 2 = 23$ degrees of freedom and multiplying the result by 2.
- For this example,

$$p_{\text{obs}} = 2\times0.00033 = 0.00066.$$

- Because p_{obs} for this example is quite small and below any commonly used confidence level (e.g., 0.01, 0.05, 0.1), we can reject the null hypothesis and conclude that blood pressure is related to sodium chloride diet level.

Example

- Data from a sample of 10 pharmacies are used to examine the relation between prescription sales volume and percentage of prescription ingredients purchased directly from the supplier.
- The sample data are shown in the following table



Pharmacy	Sales Volume, *y* (in $1,000s)	% of Ingredients Purchased Directly, *x*
1	25	10
2	55	18
3	50	25
4	75	40
5	110	50
6	138	63
7	90	42
8	60	30
9	10	5
10	100	55

a. Find the least-squares estimates for the regression line $\hat{y} = \alpha + \beta x$

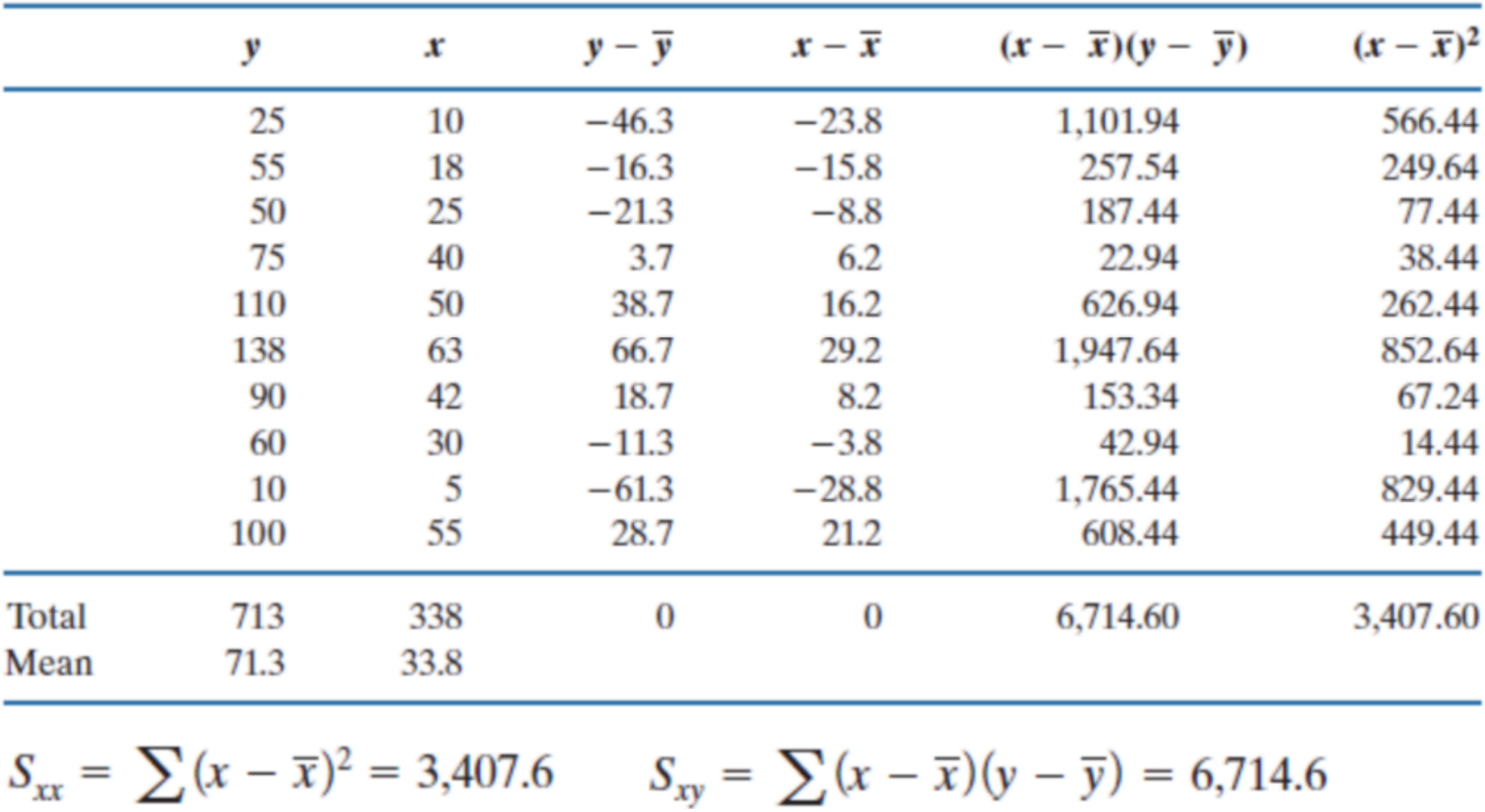
b. Predict sales volume for a pharmacy that purchases 15% of its prescription ingredients directly from the supplier.

c. Plot the (x, y) data and the prediction equation $\hat{y} = \alpha + \beta x$

d. Interpret the value of β in the context of the problem.

Example

a. Least-squares estimates



	y	x	$y - \bar{y}$	$x - \bar{x}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$
	25	10	−46.3	−23.8	1,101.94	566.44
	55	18	−16.3	−15.8	257.54	249.64
	50	25	−21.3	−8.8	187.44	77.44
	75	40	3.7	6.2	22.94	38.44
	110	50	38.7	16.2	626.94	262.44
	138	63	66.7	29.2	1,947.64	852.64
	90	42	18.7	8.2	153.34	67.24
	60	30	−11.3	−3.8	42.94	14.44
	10	5	−61.3	−28.8	1,765.44	829.44
	100	55	28.7	21.2	608.44	449.44
Total	713	338	0	0	6,714.60	3,407.60
Mean	71.3	33.8				

$$S_{xx} = \sum (x - \bar{x})^2 = 3,407.6 \qquad S_{xy} = \sum (x - \bar{x})(y - \bar{y}) = 6,714.6$$

Example

Substituting into the formulas for α and β

$$\beta = \frac{s_{xy}}{s_{xx}} = \frac{6714.6}{3407.6} = 1.97$$

$$\alpha = \bar{y} - \beta\bar{x} = 71.3 - 1.97\times33.8 = 4.7$$

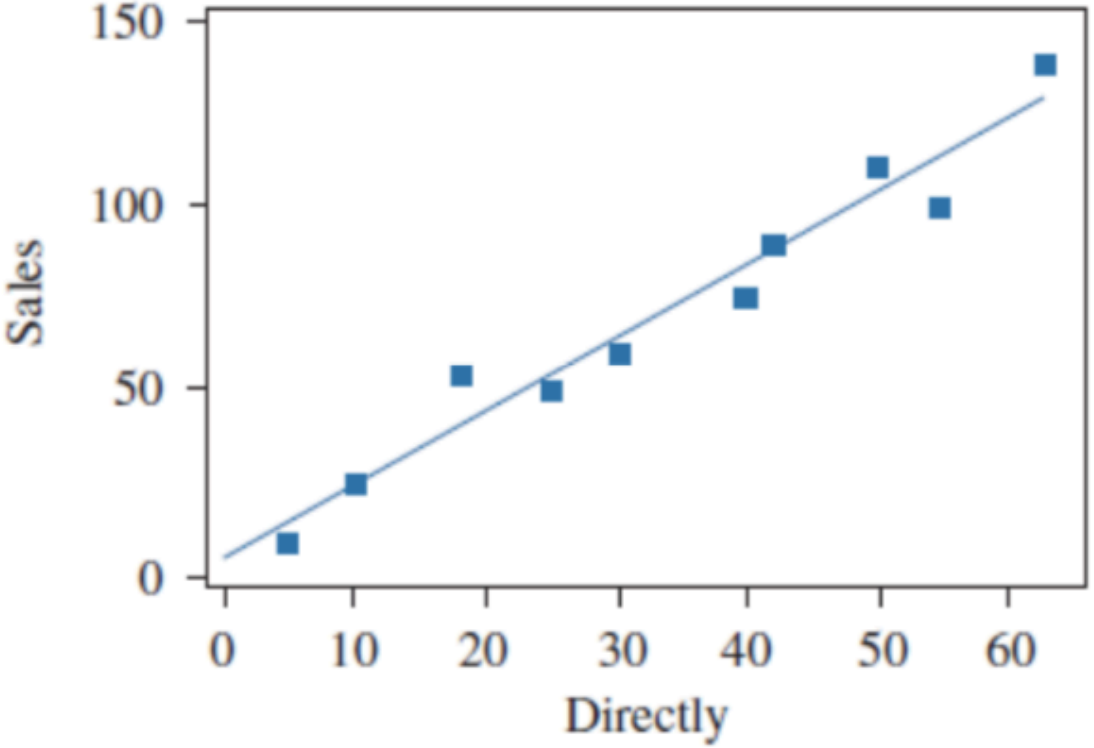
b. When x = 15%, the predicted sales volume is

$$\hat{y} = 4.7 + 1.97\times15 = 34.25$$

(that is, $34,250).

c. The (x, y) data and prediction line are plotted in the next slide:

Example



d. From $\beta = 1.97$, we conclude that if a pharmacy would increase by 1% the percentage of ingredients purchased directly, then the estimated increase in average sales volume would be \$1,970.