MAT2377

Catalin Rada

Version 2009/06/16

Hypothesis Testing

Hypothesis testing is a procedure that leads us to decide if experimental data supports a hypothesis concerning population(s) parameter(s). We will consider hypotheses concerning a population mean μ or a population proportion p. What is a hypothesis? Just a statement about the parameters of one or several populations.

Stating the Hypotheses: Often the researcher would like to verify a change in the unknown parameter under new experimental conditions. For example, a manufacturer of a new fiberglass tire claims that the mean life of the new tires is greater than the mean life of tires using the old manufacturing process. The previous mean life was $65,000$ km.

Let μ denote the mean life of the new tires. The no change hypothesis (that we will call the **null hypothesis**) is $H_0 : \mu = 65,000$ and the claim or research hypothesis (that we will call the **alternative hypothesis**) is $H_1 : \mu > 65,000$.

We want to test: $H_0 : \mu = 65,000$ against $H_1 : \mu > 65,000$.

Now we consider an example involving a proportion. Suppose that we would like to test the hypothesis that the proportion of defective items produced at a particular plant is $p = 2\%$. Then, we would test:

$$H_0 : p = 0.02 \text{ against } H_1 : p \neq 0.02.$$

Null Hypothesis: The null hypothesis will always be a simple statement concerning the unknown parameter θ. That is, it is a statement of the form $\theta = \theta_0$, where θ_0 is some real number. For example, $H_0 : \mu = 65,000$ or $H_0 : p = 0.02$. The value of the parameter in the null hypothesis will be the boundary value of the parameter from the alternative hypothesis.

Alternative Hypothesis: The alternative hypothesis will be a composite statement concerning θ. It is often the research hypothesis, i.e. the hypothesis that we would like to support with the data. We will consider three types of alternatives : (θ is the unknown parameter and θ_0 is some real number)

$H_1 : \theta < \theta_0$ is a left-sided alternative;

$H_1 : \theta > \theta_0$ is a right-sided alternative;

$H_1 : \theta \neq \theta_0$ is a two-sided alternative.

Collecting Evidence: We select a random sample of n observations and compute a point estimate for the unknown parameter θ.

Example:[Tire Example] We collect a random sample of $n = 45$ of the new fiberglass tires and observe a lifetime of $65, 158.7$ km. This is a point estimate for the true mean lifetime of such tires. Note that this evidence is in favour of the alternative hypothesis that $\mu > 65,000$. However we should not yet state that the data supports H_1.

Suppose that we decide to say that the data support $H_1 : \mu > 65,000$ if $\overline{x} > 65,000$. Now suppose that $H_0 : \mu = 65,000$ is true. What are our chances that we will say that the data support H_1? Well

$$P(\overline{X} > 65,000) = P(Z > 0) = 1 - \Phi(0) = .5$$

So there is a 50% chance that we say that the data support H_1 when in fact H_0 is true. We need to come up with a way to properly analyze the evidence.

Definitions: — A **test statistic** is a statistic that is used to test hypotheses.

— The **critical region** of the test statistic is a set of possible values of the test statistic such that if the observed value of the test statistic falls in the critical region we will reject H_0 and accept H_1.

— If we reject H_0 when H_0 is true, we say that we have committed an error of **type I** and we define:

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true})$$

— If the observed value of the test statistic does not fall in the critical region, then we fail to reject H_0. If we fail to reject H_0 when H_0 is false, then we say that we have committed an **error of type II** and

$$\beta = P(\text{type II error}) = P(\text{fail to reject } H_0 \text{ when } H_0 \text{ is false})$$

Example 1: [Tire Example Continued] Suppose that the population standard deviation is $\sigma = 1000$ km. Suppose that we use $\overline{X}$ as a test statistic.

(a) For the following critical region: $\overline{x} > 65,400$

(i) Compute the probability of committing an error of type I.

(ii) If the true mean life is $\mu = 66,000$, then compute the probability of committing an error of type II.

(iii) If the true mean life is $\mu = 67,000$, then compute the probability of committing an error of type II.

(b) For the following critical region: $\overline{x} > 65,750$.

(i) Compute the probability of committing an error of type I.

(ii) If the true mean life is $\mu = 66,000$, then compute the probability of committing an error of type II.

(iii) If the true mean life is $\mu = 67,000$, then compute the probability of committing an error of type II.

SOL: $\sigma = 1000$, and the test statistic is $\overline{X}$.

a) $\alpha = P(\text{type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true}) = P(\overline{X} > 65400; \mu = 65000) = P(\frac{\overline{X}-65000}{1000} > \frac{65400-65000}{1000}) = P(Z > 0.4) = 1 - P(Z \leq 0.4) = 1 - 0.655422 = 0.344578$

ii) $\beta = P(\text{type II error}) = P(\text{fail to reject } H_0 \text{ when } H_0 \text{ is false}) = P(\overline{X} \leq 65400, \mu = 66000) = P(\frac{\overline{X}-66000}{1000} < \frac{65400-66000}{1000}) = P(Z < -0.6) = 0.245097;$

iii) $\beta = P(\text{type II error}) = P(\text{fail to reject } H_0 \text{ when } H_0 \text{ is false}) = P(\overline{X} \leq 65400, \mu = 67000) = P(\frac{\overline{X}-67000}{1000} < \frac{65400-67000}{1000}) = P(Z < -1.6) = 0.045514$ — compare with ii).

b) $\alpha = P(\text{type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true}) = P(\overline{X} > 65750; \mu = 65000) = P(\frac{\overline{X}-65000}{1000} > \frac{65750-65000}{1000}) = P(Z > 0.75) = 1 - P(Z \leq 0.75) = 1 - 0.773373 = 0.226627.$

ii) $\beta = P(\text{type II error}) = P(\text{fail to reject } H_0 \text{ when } H_0 \text{ is false}) =$

$P(\overline{X} \leq 65750, \mu = 66000) = P(\frac{\overline{X}-66000}{1000} < \frac{65750-66000}{1000}) = P(Z < -0.25) = 0.401294;$

iii) $\beta = P(\text{type II error}) = P(\text{fail to reject } H_0 \text{ when } H_0 \text{ is false}) = P(\overline{X} \leq 65750, \mu = 67000) = P(\frac{\overline{X}-67000}{1000} < \frac{65750-67000}{1000}) = P(Z < -1.25) = 0.105650$ — compare with a).

Remarks:

— H_1 is a composite statement, i.e. a set of values. Hence β is a function of μ_1 (i.e. a possible value of μ in the alternative). If the true value of μ is close to μ_0, then β will be large, and if the true value of μ is far from μ_0, then β will be small. Since we do not know the true value of μ when H_1 is true, we cannot know the probability of making an error of type II. So if the observed value of the test statistic does not fall in the critical region, we say that we fail to reject H_0 (and never say that we accept H_0) since we cannot evaluate our chances of being wrong.

— Since $H_0 : \theta = \theta_0$ is a simple statement, we can compute α. So if the observed value falls in the critical region and we reject H_0, we say that we reject H_0 and accept H_1, since we can compute our chances of being wrong in this case.

— As we increase (decrease) our chances of making an error of type I, we increase (decrease) our chances of making an error of type II (n does not change)

— Taking these remarks into consideration, in practice, we fix the probability of making an error of type I, i.e. α. We then call, α the **level of significance** of the test. Furthermore, α should be small, but not too small, some common values are $\alpha = 10\%$, $\alpha = 5\%$ or $\alpha = 1\%$.

Test Statistic for a hypothesis concerning μ: It will be easier to construct a critical region when using a standardized test statistic. Depending on the experimental conditions we will use one of three test statistics.

(i) **Conditions:** — the population is normal or $n \geq 30$;

— σ is known.

We will use the following test statistic:

$Z_0 = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$ which follows a N(0,1) distribution if $\mu = \mu_0$.

(ii)**Conditions:** — $n \geq 40$;

— σ is unknown.

We will use the following statistic

$Z_0 = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$ which follows a N(0, 1) distribution if $\mu = \mu_0$.

(iii) **Conditions:**

— the population is normal;

— σ is unknown.

We will use the following statistic $T_0 = \frac{\overline{X}-\mu_0}{S/\sqrt{n}}$ which follows a t distribution with $\nu = n-1$ degrees of freedom if $\mu = \mu_0$ (i.e., if $H_0 : \mu = \mu_0$ is true)

Critical Region: Suppose that the test statistic is Z_0 which follows a standard normal distribution under H_0. Let z_0 be the observed value of Z_0.

(i) **[Right-Sided Alternative]** $H_1 : \mu > \mu_0$. The critical region is

$$z_0 > z_\alpha.$$

(ii) **[Left-Sided Alternative]** $H_1 : \mu < \mu_0$. The critical region is

$$z_0 < -z_\alpha.$$

(iii) **[Two-Sided Alternative]** $H_1 : \mu \neq \mu_0$. The critical region is

$$z_0 < -z_{\alpha/2} \text{ or } z_0 > z_{\alpha/2}.$$

Critical Region: Suppose that the test statistic is T_0 which follows a t distribution with $\nu = n - 1$ degrees of freedom under H_0. Let t_0 be the observed value of T_0.

(i) **[Right-Sided Alternative]** $H_1 : \mu > \mu_0$. The critical region is

$$t_0 > t_{\alpha,n-1}.$$

(ii) **[Left-Sided Alternative]** $H_1 : \mu < \mu_0$. The critical region is

$$t_0 < -t_{\alpha,n-1}.$$

(iii) **[Two-Sided Alternative]** $H_1 : \mu \neq \mu_0$. The critical region is

$$t_0 < -t_{\alpha/2,n-1} \text{ or } t_0 > t_{\alpha/2,n-1}.$$

Example 2: We would like to test the claim that the mean life of the new fiberglass tires is greater than $65,000$ km at a level of significance of 5%. A sample of $n = 45$ tires yielded a mean of 65158.7 km and a standard deviation of 1120.5 km. What are the conclusions of the test ?

Sol: From the statement we get $H_0 : \mu = 65000$, $H_1 : \mu > 65000$, $\alpha = 0.05$, $n = 45$, $\overline{x} = 65158.7$, $\sigma = 1120.5$. We compute $z_0 = \frac{\overline{x}-65000}{1120.5/\sqrt{45}} = \frac{65158.7-65000}{1120.5/\sqrt{45}} = 0.9501$ and we know that $z_\alpha = z_{0.05} = 1.645$. We reject H_0 if $z_0 > z_\alpha$, BUT that is NOT the case since 0.9501 is not greater than 1.645. So we say: we fail to reject H_0.

Example 3: A company manufactures 6-meter tubes. We randomly selected $n = 10$ tubes and computed $\overline{x} = 5.7$m and $s = 0.2$m. Can we conclude that the mean population length is not 6m at a level of significance of 5%? Assume that the population is normally distributed.

Sol: n IS NoT bigger than 40, population IS normal; σ IS unknown!

We get from the statement $H_0 : \mu = 6$, $H_1 : \mu \neq 6$, $\alpha = 0.05$, $n = 10$, $\overline{x} = 5.7$, $s = 0.2$.

What they asked? Can we reject H_0, and accept H_1 at that given level of significance? Let us see! We compute first $t_0 = \frac{\overline{x}-6}{s/\sqrt{10}} = \frac{-0.3}{0.0632} = -4.7468$, and we have $t_{\alpha/2,n-1} = t_{0.025,9} = 2.262$. Since $t_0 < -t_{\alpha/2,n-1}$ (indeed: $-4.7468 < -2.262$) we reject H_0 and accept H_1 with 0.05 level of signifiance!

P-value Method: The modern approach to hypothesis testing is to use a p-value instead of a critical region. The p-value is the probability that the test statistic will take on a value that is at least as extreme as the observed value of the statistic when the null hypothesis H_0 is true.

Def: The p-value is the smallest level of signifiance that would lead to rejection of the null hypothesis H_0 with the given data.

We compute the p-value as follows:

— If the test statistic is Z_0 which follows a N(0,1) when H_0 is true, then the p-value of the test is:

$$p = \begin{cases} 2[1 - \Phi(|z_0|)] & \text{for a two-sided alternative,} \\ 1 - \Phi(z_0) & \text{for a right-sided alternative,} \\ \Phi(z_0) & \text{for a left-sided alternative,} \end{cases}$$

— If the test statistic is T_0 which follows a t distribution with $\nu = n - 1$ degrees of freedom when H_0 is true, let T be a t random variable with $n - 1$ degrees of freedom. Then the p-value of the test is:

$$p = \begin{cases} 2P(T > |t_0|) & \text{for a two-sided alternative,} \\ P(T > t_0) & \text{for a right-sided alternative,} \\ P(T < t_0) & \text{for a left-sided alternative,} \end{cases}$$

Decision Rule: $p < \alpha$ is equivalent to the observed value falling in the critical region. Hence, we will reject H_0 if $p < \alpha$.

Example 4: Answer Exp 2 using the p-value method.

Sol: Same data as in Exp 2, so take from there $z_0 = 0.9501$. Compute $p = 1 - \Phi(0.9501) \cong 1 - \Phi(0.95) = 1 - 0.828944 = 0.171056$. Comparison: since $p = 0.171056$ IS not $<$ than $\alpha = 0.05$, we conlcude that we fail to reject H_0 (as in the previous solution).

Example 5: Answer Exp 3 using the p-value method.

Sol: Same data as in Exp 3, so take from there $t_0 = -4.7468$. We compute $p = 2P(T > |t_0|) = 2P(T > |-4.7468|) = 2P(T > 4.7468) = 2P(T_{\alpha,n-1} > 4.7468) = 2P(T_{0.05,9} > 4.7468) = 2 \times 0.0005 = 0.001$. Comparison: since $p = 0.001 < \alpha = 0.05$ we reject H_0, as in the previous solution!

Type II error and Choice of Sample Size:

We will only discuss type II error and the choice of sample size for a normal population with σ known.

Suppose that we would like to find the probability of making a type II error (recall β) when we assume that the true mean is μ_1. In other words: $H_0 : \mu = \mu_0$; $H_1 : \mu \neq \mu_0$. Define $\delta = \mu_1 - \mu_0$. The test statistic $Z_0 = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} = \frac{\overline{X}+\delta-\mu_1}{\sigma/\sqrt{n}} = \frac{\overline{X}-\mu_1}{\sigma/\sqrt{n}} + \frac{\delta}{\sigma/\sqrt{n}}$. Since the population is normal (and the true mean is μ_1) we get that $\overline{X}$ follows a $N(\mu_1, \sigma^2/n)$, and thus (by the old tricks) Z_0 follows a $N(\frac{\delta\sqrt{n}}{\sigma}, 1)$ distribution.

How do we compute β? Recall the critical regions and so $\beta =P(\text{type II error}) = P(-z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2}) = P(-z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma} \leq Z_0 - \frac{\delta\sqrt{n}}{\sigma} \leq z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}) = P(-z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma} \leq Z \leq z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}) = \Phi(z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}) - \Phi(-z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma})$, where Z is standard normal...

Example 6: Suppose that we are testing $H_0 : \mu = 5$ against $H_1 : \mu \neq 5$ at $\alpha = 5\%$. The population is normal with $\sigma = 0.5$. If the sample size is $n = 10$, what is the probability of committing an error of type II when the true mean is 6?

sol: $\beta = \Phi(z_{0.025} - \frac{(6-5)\sqrt{10}}{0.5}) - \Phi(-z_{0.025} - \frac{(6-5)\sqrt{10}}{0.5}) = etc$ (use $z_{0.025} = 1.96$)

Sample Size:

To control the probability of error of type II, that is to obtain a particular value of β for a given α and δ, at $\mu_1 = \mu_0 + \delta$, we can choose an appropriate sample size:

For a two-sided alternative: We require the following sample size:

$$n \approx \frac{(z_{\alpha/2}+z_{\beta})^2\sigma^2}{\delta^2},$$

For a one-sided alternative: We require the following sample size

$$n = \frac{(z_{\alpha}+z_{\beta})^2\sigma^2}{\delta^2}.$$

Example 7: Suppose that we are testing $H_0 : \mu = 5$ against $H_1 : \mu \neq 5$ at $\alpha = 5\%$. The population is normal with $\sigma = 0.5$. We want to design the experiment in such way that if $\mu = 6$, then we want the probability of failing to reject H_0 to be 2%. Determine the required sample size.

Sol: by the definition of β we have $\beta = 0.02$. So $n \approx \frac{(z_{\alpha/2}+z_\beta)^2\sigma^2}{\delta^2} = \frac{(z_{0.01}+z_{0.02})^2(0.5)^2}{(6-5)^2} = (2.32 + 2.055)^2 \times (0.5)^2 = 4.7851$, so we round up $n = 5$.

9-5 Hypothesis testing concerning a proportion p

Suppose that we want to test the null hypothesis $H_0 : p = p_0$, where p is an unknown population proportion. We will use a test statistic based upon the sample proportion $\hat{P}$. (Recall that $\hat{P} = \frac{X}{n}$) The test statistic (in this case) is

$Z_0 = \frac{\hat{P}-p_0}{\sqrt{p_0(1-p_0)/n}}$, and it follows approximately a $N(0,1)$ distribution when H_0 is true and $np_0 \geq 5$ and $n(1-p_0) \geq 5$.

Critical Region: Let z_0 be the observed value of Z_0.

(i) **[Right-Sided Alternative]** $H_1 : p > p_0$. The critical region is $z_0 > z_\alpha$.

(ii) **[Left-Sided Alternative]** $H_1 : p < p_0$. The critical region is $z_0 < -z_\alpha$.

(iii) **[Two-Sided Alternative]** $H_1 : p \neq p_0$. The critical region is $z_0 < -z_{\alpha/2}$ or $z_0 > z_{\alpha/2}$.

Example 8: Suppose that we would like to test the hypothesis that the proportion of defective items produced at a particular plant is $p = 2\%$. From $n = 500$ random selected items there are 8 which are defective. Do the data suggest that $p \neq .02$ at $\alpha = 5\%$?

Sol: from the statement we see that we are dealing with a 2-sided alternative! We have $H_0 : p = 0.02$, $H_1 : p \neq 0.02$, $\alpha = 0.05$, so $\alpha/2 = 0.025$; $n = 500$, $X = 8$, so $\hat{P} = \frac{X}{n} = \frac{8}{500}$. So we get $z_0 =$

$\frac{\frac{8}{500}-0.02}{\sqrt{0.02(1-0.02)/500}} = \frac{-0.004}{\sqrt{0.0000392}} \cong -0.6388766$. Recall that $z_{\alpha/2} = z_{0.025} = 1.96$.

Question? Is our observed value z_0 in $(-\infty, -1.96)$ or in $(1.96, \infty)$? NO! So the observed is not the critical region, hence we fail to reject H_0.

Sample Size:

To control the probability of error of type II, that is to obtain a particular value of β for a given α, using the same type of ideas as in the previous lecture one can choose an appropriate sample size (where $H_1 : p \neq p_0$, and p is the true value of the population proportion):

$$n = \left[\frac{z_{\alpha/2}\sqrt{p_0(1-p_0)}+z_\beta\sqrt{p(1-p)}}{p-p_0}\right]^2$$

If the alternative is a one sided alternative, then $n =$

$\left[\frac{z_\alpha\sqrt{p_0(1-p_0)}+z_\beta\sqrt{p(1-p)}}{p-p_0}\right]^2$

Exp 9: (see page 330) Assume that 500 beers are tested and 10 are rejected. Test the hypothesis $H_0 : p = 0.03$ against $H_1 : p < 0.03$ at $\alpha = 0.05$.

Sol: We are interested in the proportion of rejected beers; $H_0 : p = 0.03$, $H_1 : p < 0.03$ (one sided alternative); $\alpha = 0.05$, $p_0 = 0.03$. Let us compute our point estimate as follows: $z_0 = \frac{\hat{P}-p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{\frac{10}{500}-0.03}{\sqrt{(0.03\times 0.97)/500}} = \frac{-0.01}{0.007629} \cong -1.31$.

Question? Is $z_0 < -z_\alpha$? OR: is $-1.31 < -z_{0.05}$? OR: is $-1.31 < -1.65$? NO! So we fail to reject H_0; there is not enough evidence at this level of signifiance!

Exp: 10 (regarding the previous lecture) Components are manufactured to have strength normally distributed with mean $\mu = 40$ units and standard deviation $\sigma = 1.2$ units. A modification has been tried, for which an increase in mean strength is claimed (the standard deviation remains the same). A random sample of $n = 12$ components produced using the modified process had strength

`42.5, 39.8, 40.3, 43.1, 39.6, 41.0, 39.9, 42.1, 40.7, 41.6, 42.1, 40.8,`

Do the data provide strong evidence that the mean strength exceeds 40 units? ($Use\ \alpha = 0.05$)

Sol: $H_0 : \mu = 40$, $H_1 : \mu > 40$, and the test statistics is $Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$ under H_0. Then we have $\bar{x} = 41.125$. Then we compute the observed value $z_0 = 3.248$. We do have a one-sided alternative, so the critical region is in this case: $z_0 > z_\alpha$. Since $\alpha = 0.05$, we get $z_\alpha = 1.65$. Now just do the comparison: since indeed $z_0 = 3.248 > 1.65 = z_\alpha$ we reject H_0 and accept H_1.

One more Exp (11) Assume we have the following data

18.0	17.4	15.5	16.8	19.0	17.8	17.4	15.8
17.9	16.3	16.9	18.6	17.7	16.4	18.2	18.7

from $\mathcal{N}(\mu, \sigma^2)$ with completely unknown μ and σ^2. Test $H_0 : \mu = 16.6$ against $H_1 : \mu > 16.6$. Use $\alpha = 0.05$.

Sol: The easy part: $H_0 : \mu = 16.6$, $H_1 : \mu > 16.6$, $\alpha = 0.05$; $n = 16$. Hard part (?): we are in the case iii)/page 10. So we use $T_0 = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim \mathcal{N}(0,1)$ under H_0. We have $\bar{x} = 17.4$ and $s^2 = 1.039$. The observed value t_0 of this statistics is $t_0 = \frac{17.4-16.6}{\sqrt{1.039}/\sqrt{16}} = 3.081$.

Since $\alpha = 0.05$ and $n = 16$, we get $t_{\alpha,n-1} = t_{0.05,15} = 1.753$ (keep in mind that you do have a one-sided alternative, so you should know that the critical region is in this case:$t_0 > t_{\alpha,n-1}$) We reject H_0 because $t_0 > t_{\alpha,n-1}$.

Joint Probability Distributions

For discrete distributions : Refer to Section 5-1; For continuous distributions : Refer to Section 5-2

Def: Suppose that we have two random variables X, Y defined on a common sample space Ω, then we say that (X, Y) is a random vector.

Discrete Joint distribution: Let X and Y be two discrete random variables. The joint probability mass function of X and Y is defined as:

$$f_{XY}(x, y) = P(X = x, Y = y) = P(X = x \text{ and } Y = y).$$

The range of the random vector (X, Y) is $R_{XY} = \{(x, y) : f_{XY}(x, y) \neq 0\}$. **Properties of the joint distribution:**

1. (non-zero probability) $f_{XY}(x, y) \geq 0$

2. (total mass =1) $\sum_{(x,y)\in R_{XY}} f_{XY}(x,y) = 1$

3. (computational property) $P((X,Y) \in A) = \sum_{(x,y)\in A\cap R_{XY}} f_{XY}(x,y)$.

Definition: If X and Y are discrete random variables with joint probability mass function f_{XY}, then the marginal probability mass functions of X and Y are respectively:

$f_X(x) = P(X = x) = \sum_y f_{XY}(x,y)$ and $f_Y(y) = P(Y = y) = \sum_x f_{XY}(x,y)$

Independence: We will say that X and Y are independent if:

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

Example 1: Consider the following joint probability mass function:

x	y	$f_{XY}(x,y)$
1	1	1/4
1.5	2	1/8
1.5	3	1/4
2.5	4	1/4
3	5	1/8

Determine the following probabilities : (a) $P(X < 2.5, Y < 3)$

(b) $P(X < 2.5)$, (c) $P(X > 1.8, Y > 4.7)$

(d) $P(Y > 2|X = 1.5)$, (e) Find the marginal distribution of X.

(f) Compute the mean of X; (g) Are X and Y independent ?

Sol: a) $P(X < 2.5, Y < 3) = P((X,Y) \in \{(1,1),(1.5,2)\}) = f_{XY}(1,1) + f_{XY}(1.5,2) = 1/4 + 1/8 = 3/8$; b) $P(X < 2.5) =$

$P(X < 2.5, Y$ any $) = P((X, Y) \in \{(1,1), (1.5,2), (1.5,3)\}) = f(1,1) + f(1.5,2) + f(1.5,3) = 1/4 + 1/8 + 1/4 = 5/8$; (c) $P(X > 1.8, Y > 4.7) = P((X,Y) = (3,5)) = 1/8$; (d) $P(Y > 2|X = 1.5) = \frac{P(Y>2 \text{ and } X=1.5)}{P(X=1.5)} = \frac{P((X,Y)=(1.5,3))}{\sum_y f(1.5,y)} = \frac{1/4}{f(1.5,2)+f(1.5,3)} = \frac{1/4}{1/8+1/4} = 2/3$;

e) $f_X(1) = P(X = 1) = \sum_y f_{XY}(1, y) = f(1,1) = 1/4$; $f_X(2.5) = P(X = 2.5) = \sum_y f_{XY}(2.5, y) = f(2.5, 4) = 1/4$; $f_X(3) = P(X = 3) = \sum_y f_{XY}(3, y) = f(3,5) = 1/8$; the rest is above...

f) $E(X) = \sum_x x f_X(x) = 1 \times \frac{1}{4} + (1.5) \times \frac{3}{8} + (2.5) \times \frac{1}{4} + 3 \times \frac{1}{8} = \ldots$

g) Note that $f_Y(1) = P(Y = 1) = P((X,Y) = (1,1)) = f_{XY}(1,1) = 1/4$, and so $f_X x f_Y y = \frac{1}{4} \times \frac{1}{4} \neq 1/4 = f_{XY}(1,1)$, therefore not independent!

Continuous Joint distribution: Let X and Y be two continuous random variables. To specify the probabilities associated with the random vector (X, Y) we can define a probability density function f_{XY} such that:

$P[(X,Y) \in R] = \int\int_R f_{XY}(x,y)dxdy$, where R is just a subset: $R \subseteq \mathbf{R}^2 = \{(x,y) : x \in \mathbf{R}$ and $y \in \mathbf{R}\}$.

Properties of the joint distribution:

1. (non-zero density) $f_{XY}(x,y) \geq 0$;
2. (total mass =1) $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{XY}(x,y) = 1$;
3. (computational property) $P[(X,Y) \in R] = \int\int_R f_{XY}(x,y)dxdy$.

Definition: If X and Y are continuous random variables with joint probability density function f_{XY}, then the marginal probability density functions of X and Y are respectively:

$f_X(x) = \int_y f_{XY}(x,y)dy$ and $f_Y(y) = \int_x f_{XY}(x,y)dx$.

Independence: We will say that X and Y are independent if:

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

Example 3: Consider the joint probability density function

$f_{XY}(x,y) = cxy$, $0 < x < 1$, $x < y < x+2$.

(a) Determine the value of the constant c.

(b) Determine $P(Y - X > 1)$.

(c) Determine the marginal probability density functions f_X and f_Y.

Sol: at the blackboard: for a) integral is 1, so $c = 3/5$; for b) do a picture; c) $f_X(x) = \int_x^{x+2} f_{XY}(x,y)dy$ and the more complicated part is:

$$f_Y(y) = \begin{cases} \int_0^y \frac{3}{5}xydx & \text{if } y \in (0,1], \\ \int_0^1 \frac{3}{5}xydx & \text{if } y \in (1,2], \\ \int_{y-2}^1 \frac{3}{5}xydx & \text{if } y \in (2,3). \end{cases}$$

Finish the computations! One may integrate both marginal p.d.f.s and one gets 1 in both cases (as we expected!). Always draw a picture when not sure about the subets of $\mathbf{R}^2$. For a): changing the order of integration implies that the limits of integration (the boundaries) are changing!

5-3 Covariance and Correlation

Goal: Try to describe the relationship between Y and X.

A common measure of the relationship between two random variables is the covariance. Before we can define the covariance we need to define the expectation of a function of two random variables.

Definition: Let X and Y be two random variables. The expectation of $h(X, Y)$ is defined by:

$E[h(X,Y)] = \sum\sum h(x,y) f_{XY}(x,y)$, if the rvs X, Y are discrete;

$E[h(X,Y)] = \int\int h(x,y) f_{XY}(x,y) dxdy$, if the rvs X, Y are continuous.

Interpretation: $E[h(X,Y)]$ is the average value of $h(X,Y)$ that is expected in a long sequence of repeated trials of the random experiment.

Example 1: Let X and Y be two discrete random variables. Show that $E[aX+bY]=aE[X]+bE[Y]$. (the same holds for continuous rvs, just replace in the proof sum by integral)

sol: at the blackboard

Example 2: Let X and Y be two random variables. Show that $V[X+Y]=V[X]+V[Y]+2E[(X-\mu_X)(Y-\mu_Y)]$.

Remark : The quantity in the last term will be used as a measure of the linear relationship between X and Y.

Sol: at blackboard.

Definition: The covariance between the random variables X and Y is

$\sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - \mu_X\mu_Y$.

Example 3: Consider the following joint distribution:

x	y	$f_{XY}(x,y)$
1	1	1/4
1.5	2	1/8
1.5	3	1/4
2.5	4	1/4
3	5	1/8

Find the covariance between X and Y.

Sol: Using f) from the previous table we get $\mu_X = 29/16$. From g) we know already $f_Y(1) = 1/4$. Similarly one gets: $f_Y(2) = 1/8$, $f_Y(3) = 1/4$, $f_Y(4) = 1/4$, $f_Y(5) = 1/8$. Hence $\mu_Y = 1\times\frac{1}{4}+2\times\frac{1}{8}+3\times\frac{1}{4}+4\times\frac{1}{4}+5\times\frac{1}{8} = 23/8$.

Looking at the table one gets: $E(XY) = 1 \times 1 \times \frac{1}{4} + 1.5 \times 2 \times \frac{1}{8} + 1.5 \times 3 \times \frac{1}{4} + 2.5 \times 4 \times \frac{1}{4} + 3 \times 5 \times \frac{1}{8} = 6.125.$

So, the covariance between X and Y is $6.125 - 29/16 \times 23/8 = 6.125 - 5.210 = 0.915.$

Remark: In practice, we often use a unit-less version of the covariance which is called the correlation coeficient. It is easier to interpret it since it can be shown to fall between -1 and 1.

Definition: The correlation coeficient between X and Y is $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_X}$, where σ_X and σ_Y are respectively the standard deviation for X and Y.

Properties of the correlation coeficient:

1. $-1 \leq \rho \leq 1$,

2. If the points taken by (X, Y) fall exactly on a line, then $\rho_{XY} = 1$ or $\rho_{XY} = -1$. The sign will be positive if the slope is positive and negative if the slope is negative.

3. If X and Y are independent, then $\rho_{XY} = \sigma_{XY} = 0$.

Note: However, $\rho_{XY} = \sigma_{XY} = 0$, does not necessarily imply the independence of X and Y.

Remarks: — If X and Y have a non-zero correlation, then we say that they are correlated and *thus not dependent*;

— If X and Y have a zero correlation, then we say that they are uncorrelated. However, we cannot say anything about independence.

Example 4: (a) Find the correlation coeficient between X and Y from example 3;

(b) Are X and Y independent ?

Sol: a) We have $\sigma_X = \sqrt{Var(X)} = \sqrt{E(X-\mu_X)^2} = \sqrt{\sum_x x^2 f_X(x) - \mu_X^2} = \sqrt{1^2 \times \frac{1}{4} + (1.5)^2 \times \frac{3}{8} + (2.5)^2 \times \frac{1}{4} + 3^2 \times \frac{1}{8} - (\frac{29}{16})^2}$ by pages 29 and 35. So $\sigma_X = 0.7043$.

We have $\sigma_Y = \sqrt{Var(Y)} = \sqrt{E(Y-\mu_Y)^2} = \sqrt{\sum_y y^2 f_Y(y) - \mu_Y^2} = \sqrt{1^2 \times \frac{1}{4} + (2)^2 \times \frac{1}{8} + (3)^2 \times \frac{1}{4} + 4^2 \times \frac{1}{4} + 5^2 \times \frac{1}{8} - (\frac{23}{8})^2}$ by page 35. So $\sigma_Y = 1.3636$.

Hence (by page): $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_X} = \frac{0.915}{0.7043 \times 1.3636} = 0.95275 \neq 0$, thus not independent, i.e., dependent, i.e., X and Y are correlated!

Example 5: Consider the joint distribution:

x	-1	0	0	1
y	0	-1	1	0
$f_{XY}(x,y)$	1/4	1/4	1/4	1/4

Show that the correlation coeficient between X and Y is zero, but X and Y are not independent.

Sol: Since $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_X}$, we compute $\sigma_{XY} = E(XY) - \mu_X \mu_Y$ as follows: $E(XY) = \sum_{x,y} xy f_{XY}(x,y) = (-1) \times 0 \times \frac{1}{4} + (0) \times -1 \times \frac{1}{4} + (0) \times 1 \times \frac{1}{4} + (1) \times 0 \times \frac{1}{4} = 0$ (wow), then $E(X) = \sum_{x,y} x f_{XY}(x,y) = (-1) \times \frac{1}{4} + (0) \times \frac{1}{4} + (0) \times \frac{1}{4} + (1) \times \frac{1}{4} = 0$, and $E(Y) = \sum_{x,y} y f_{XY}(x,y) = (0) \times \frac{1}{4} + (-1) \times \frac{1}{4} + (1) \times \frac{1}{4} + (0) \times \frac{1}{4} = 0$. We get that $\sigma_{XY} = E(XY) - \mu_X \mu_Y = 0$, so $\rho_{XY} = 0$.

For the second part note that $f_X(-1) = \sum_y f_{XY}(-1,y) = f(-1,0) = 1/4$ and that $f_Y(0) = \sum_x f_{XY}(x,0) = f(-1,0) + f(1,0) = \frac{1}{4} + \frac{1}{4} = 1/2$.

Since $f_X(-1)f_Y(0) = \frac{1}{4} \times \frac{1}{2} \neq 1/4 = f_{XY}(-1,0)$ we get that X and Y are not independent.