Tests of Hypotheses Based on a Single Sample (Devore Chapter Eight)

MATH-252-01: Probability and Statistics II[*]

Fall 2016

Contents

Thursday 15 September 2016

1 Overview of Hypothesis Testing

We now consider another area of statistical inference known as hypothesis testing. The usual formulation starts with a null hypothesis H_0 and an alternative hypothesis H_a, which produce different probabilistic predictions about the outcome of an experiment, and then, based on the observed data, decides between two alternatives:

1. Reject H_0
2. Don't reject H_0

The full scope of hypothesis testing is quite general, but for this introduction, we'll make some simplifying assumptions:

1. The data $\{X_i\}$ are a sample of size n from a probability distribution with pdf $f(x;\theta)$ (or pmf $p(x;\theta)$, if it's a discrete distribution).
2. The null hypothesis H_0 specifies a single value for the parameter $\theta = \theta_0$. (This is known as a "point hypothesis" because it gives a single value of θ completely specifies the distribution.)

3. The alternative hypothesis H_a specifies some range of values for θ which are inconsistent with θ_0, typically one of the following:
   (a) $\theta \neq \theta_0$
   (b) $\theta > \theta_0$
   (c) $\theta < \theta_0$

   (Any of these is a "composite hypothesis" because it corresponds to a set of θ values, and therefore to a family of distributions.)
4. The test is defined by constructing a statistic $Y = u(X_1, \ldots, X_n)$ and rejecting H_0 or not according to the value of Y.

For example, suppose we are testing the claims of a psychic who is allegedly able to determine the suit of a card drawn from a poker deck (clubs, diamonds, hearts, spades) without seeing it, but whose ESP is imperfect. If we reshuffle the deck after each draw, the test statistic for n draws is a binomial random variable $Y \sim \text{Bin}(n, \theta)$. The null hypothesis H_0 is $\theta = 0.25$ (no ESP; random guesses out of four suits), and the alternative hypothesis H_a is $\theta > 0.25$ (we assume a psychic will do *better* than random guessing, not worse). Suppose we do 20 trials; one test we could use is to reject H_0 if the psychic gets more than 8 of them correct. (We'd expect 5 from pure guessing.) So if the number of correct answers is 9 or more, we reject H_0, but if it's 8 or fewer, we do not. We call $Y > 8$ the *rejection region* for this test.

1.1 Type I and Type II Errors

Because of the random nature of the experiment, there will be some probability that the test will reject H_0. Even if the null hypothesis H_0 is true, we will generally have a non-zero probability of rejecting it. Likewise, even if the alternative hypothesis H_a is true, the probability that the data will lead us to reject H_0 will still generally be less than one. A perfect test would have us never reject H_0 if it's true, and always reject H_0 if H_a is true, but in most situations there is no perfect test. A given test thus has some probability of making an error. If H_0 is true and we reject it. this is called a *Type I Error*, also known as a false alarm. (We have claimed to see an effect which was not there.) If H_a is true, but we do not reject H_0, this is called a *Type II Error*, also known as a false dismissal. (We have failed to find an effect which is there.) The probability of each of these errors happening has to be understood as a conditional probability (since it assumes one hypothesis or the other is true). The probability of a type I error, or the false alarm probability, is

$$\alpha = P(\text{reject } H_0 | H_0 \text{ is true}) \tag{1.1}$$

The probability of a type II error, or the false dismissal probability, is

$$\beta = 1 - P(\text{reject } H_0 | H_a \text{ is true}) \tag{1.2}$$

Actually, since we're taking the alternative hypothesis H_a to be a composite hypothesis, this depends on the actual value of θ:

$$\beta(\theta) = 1 - P(\text{reject } H_0 | \text{parameter value } \theta) \tag{1.3}$$

We'd generally like to have α and β as small as possible. In practice, one usually decides what false alarm probability α one can afford, and then designs a test which minimizes $\beta(\theta)$ for any θ given that constraint.

A related quantity is the *power* of the test, which is the probability of rejecting H_0 if H_a is true. It is written $\gamma(\theta)$ and equal to $1 - \beta(\theta)$.

For the case of the ESP test described above, the false alarm and false dismissal probabilities can be determined from the binomial cdf:

$$\begin{aligned}\alpha = P(Y > 8|\theta = 0.25) &= \sum_{x=9}^{20} \binom{20}{x} 0.25^x 0.75^{20-x} \\ &= 1 - F(8; 20, 0.25) \approx 1 - .959 = 0.041\end{aligned} \tag{1.4}$$

where we've looked up the binomial cdf from Table A.1 in the back of Devore. Likewise, if the true θ is 0.50, the false dismissal probability of the test will be

$$\beta(0.20) = P(Y \le 8|\theta = 0.50) = F(8; 20, 0.50) \approx .252 \tag{1.5}$$

2 Tests Concerning the Mean

2.1 Choice of Exclusion Region

Suppose you have a sample of size n drawn from some distribution, and a null hypothesis H_0 that the mean of this distribution is $\mu = \mu_0$. You're given a tolerable false-alarm rate α (say 5% or 10%) and want to construct a test of H_0 given with that α. We can construct this using the same statistics that we used to make confidence intervals. In particular

1. If the sample is drawn from a normal distribution (or if $n \gtrsim 30$, almost any distribution) with a known variance σ^2, the statistic $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ obeys a standard normal distribution, if H_0 is true. Here $\overline{X} = \sum_{i=1}^{n} X_i$ is the sample mean.
2. If we have a large sample ($n \gtrsim 40$) from a distribution (normal or otherwise) with a finite but unknown variance, the statistic $Z = \frac{\overline{X}-\mu_0}{\sqrt{S^2/n}}$ is approximately standard-normal distributed, if H_0 is true. Here $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is the sample variance.
3. If the sample is drawn from a normal distribution with unknown variance, the statistic $T = \frac{\overline{X}-\mu_0}{\sqrt{S^2/n}}$ obeys a Student t-distribution with $n-1$ degrees of freedom.

The construction of the tests thus uses the percentiles of either the standard normal or Student t distribution as appropriate.

We'll look in detail at the first case, since the construction in the other cases is analogous. First, the question is whether to reject H_0 if the statistic $\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ is "too high", "too low", or both. This depends on the form of the alternative hypothesis H_a. (In fact, this choice of direction is basically the only way that the alternative hypothesis affects the construction of the test.) The probabilistic statements of interest are

$$P\left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} > z_\alpha \middle| \mu = \mu_0\right) = \alpha \tag{2.1a}$$

$$P\left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} < -z_\alpha \middle| \mu = \mu_0\right) = \alpha \tag{2.1b}$$

and

$$P\left(\left[\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} < -z_{\alpha/2}\right] \bigcup \left[\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2}\right] \middle| \mu = \mu_0\right) = \alpha \tag{2.1c}$$

Any one of these can be used to create a rejection region with false alarm probability α. Which one you want depends on the alternative hypothesis. If H_a is $\mu > \mu_0$, you want to reject H_0 if $\overline{X}$ is a lot more than μ_0, i.e., if Z is too large. (If $\overline{X}$ is a lot less than μ_0, then the null hypothesis H_0 is a bad fit to the data, but the alternative hypothesis H_a is even worse.) So in that case we use $Z > z_\alpha$ as our rejection region. Similarly, if H_a is $\mu < \mu_0$, we reject H_0 if $\overline{X}$ is too far below μ_0, i.e., if $Z < z_\alpha$. If H_a is $\mu \neq \mu_0$, then we $\overline{X}$ either too far above or below μ_0 would be an inconsistency with H_0 which was more consistent with H_a. So, to summarize, **for a normal distribution with known variance σ^2**:

1. If H_a is $\mu > \mu_0$, we reject H_0 if $\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} > z_\alpha$. This is called an upper-tailed test.
2. If H_a is $\mu < \mu_0$, we reject H_0 if $\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} < -z_\alpha$. This is called a lower-tailed test.
3. If H_a is $\mu \neq \mu_0$, we reject H_0 if either $\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} < -z_{\alpha/2}$ or $\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2}$. This is called a two-tailed test.

Similarly, **for a large sample from any distribution with unknown σ**:

1. If H_a is $\mu > \mu_0$, we reject H_0 if $\frac{\overline{X}-\mu_0}{\sqrt{S^2/n}} > z_\alpha$.
2. If H_a is $\mu < \mu_0$, we reject H_0 if $\frac{\overline{X}-\mu_0}{\sqrt{S^2/n}} < -z_\alpha$.
3. If H_a is $\mu \neq \mu_0$, we reject H_0 if either $\frac{\overline{X}-\mu_0}{\sqrt{S^2/n}} < -z_{\alpha/2}$ or $\frac{\overline{X}-\mu_0}{\sqrt{S^2/n}} > z_{\alpha/2}$.

and finally, **for a sample from a normal distribution with unknown σ**:

1. If H_a is $\mu > \mu_0$, we reject H_0 if $\frac{\overline{X}-\mu_0}{\sqrt{S^2/n}} > t_{\alpha;n-1}$.
2. If H_a is $\mu < \mu_0$, we reject H_0 if $\frac{\overline{X}-\mu_0}{\sqrt{S^2/n}} < -t_{\alpha;n-1}$.
3. If H_a is $\mu \neq \mu_0$, we reject H_0 if either $\frac{\overline{X}-\mu_0}{\sqrt{S^2/n}} < -t_{\alpha/2;n-1}$ or $\frac{\overline{X}-\mu_0}{\sqrt{S^2/n}} > t_{\alpha/2;n-1}$.

The two sorts of tests using the standard normal percentiles are called z tests; the one using the Student t percentiles is called a t test.

Practice Problems

8.1, 8.7, 8.9, 8.13, 8.19, 8.31

Tuesday 20 September 2016

2.2 False Dismissal Probability

We limit attention to the first case, which is the most straightforward.

To get the false dismissal probability $\beta(\mu)$, or equivalently the power $\gamma(\mu) = 1 - \beta(\mu)$, we need to consider the probability of the sample landing in the rejection region for a given $\mu = \mu'$ consisted with the alternative hypothesis H_a. In the case of a normal distribution with known σ, the test statistic $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ will still be normally distributed, but now, since $\overline{X} \sim (\mu', \sigma^2/n)$, the mean of Z will be $\frac{\mu'-\mu_0}{\sigma/\sqrt{n}}$. (The variance will still be 1.) Thus, if H_a is $\mu > \mu_0$

$$\beta(\mu') = P\left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \leq z_\alpha \middle| \mu = \mu'\right) = \Phi\left(z_\alpha - \frac{\mu'-\mu_0}{\sigma/\sqrt{n}}\right) \tag{2.2}$$

while if H_a is $\mu < \mu_0$

$$\begin{aligned}\beta(\mu') &= P\left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \geq -z_\alpha \middle| \mu = \mu'\right) = 1 - \Phi\left(-z_\alpha - \frac{\mu'-\mu_0}{\sigma/\sqrt{n}}\right) \\ &= \Phi\left(z_\alpha - \frac{\mu_0-\mu'}{\sigma/\sqrt{n}}\right)\end{aligned} \tag{2.3}$$

Note that $\Phi(z_\alpha) = 1 - \alpha$, and in each case the argument is less than z_α, so $\beta(\mu') < 1 - \alpha$, which means $\gamma(\mu') > \alpha$. This makes sense, since you'd expect the test to be more likely to reject H_0 if H_a is true than if H_0 is true.

For a two-tailed test, the calculation of the false dismissal probability is also straightforward:

$$\begin{aligned}\beta(\mu') &= P\left(-z_{\alpha/2} \leq \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \middle| \mu=\mu'\right) \\ &= \Phi\left(z_{\alpha/2} - \frac{\mu'-\mu_0}{\sigma/\sqrt{n}}\right) - \Phi\left(-z_{\alpha/2} - \frac{\mu'-\mu_0}{\sigma/\sqrt{n}}\right)\end{aligned} \tag{2.4}$$

2.3 Sample Size Determination

We can turn the false dismissal probability expressions for one-tailed tests around, and ask what sample size n will allow us to produce a test with a specified false alarm probability α and false dismissal probability β for a nominal population mean μ'. We use the fact that

$$\beta = 1 - \Phi(z_\beta) = \Phi(-z_\beta) \ , \tag{2.5}$$

which means that

$$-z_\beta = \begin{cases} z_\alpha - \frac{\mu'-\mu_0}{\sigma/\sqrt{n}} & \text{upper tailed} \\ z_\alpha - \frac{\mu_0-\mu'}{\sigma/\sqrt{n}} & \text{lower tailed} \end{cases} \tag{2.6}$$

In either case if we solve for n we get the minimum sample size

$$n = \left(\frac{\sigma(z_\alpha + z_\beta)}{\mu'-\mu_0}\right)^2 \tag{2.7}$$

3 Tests Concerning Proportion

Now we turn once again to the case of a binomial-type experiment, e.g., sampling n members from a large population where some fraction (or proportion) p of the members have some desired trait, or doing n independent trials with a probability p for success on each trial. As usual, the language about sample and random variables is a little different. We could consider this to be a sample of size n from a Bernoulli distribution $\text{Bin}(1,p)$, or a single binomial random variable $X \sim \text{Bin}(n,p)$. In any event, it's more convenient to work with the estimator $\hat{p} = X/n$, which has mean

$$E(\hat{p}) = \frac{np}{n} = p \tag{3.1}$$

and variance

$$V(\hat{p}) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n} \tag{3.2}$$

We can consider two regimes when testing a null hypothesis H_0 which states $p = p_0$: if $np_0 \gtrsim 10$ and $n(1-p_0) \gtrsim 10$, we can treat the distribution of $\hat{p}$ as approximately normal with the mean and variance given above, which means we can use the testing procedures already defined. If not, we need to use the binomial cumulative distribution function to define and evaluate the tests.

3.1 Large Sample Tests

Supposing the normal approximation to be valid, the test statistic appropriate when the null hypothesis H_0 is $p = p_0$ will be

$$Z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} \tag{3.3}$$

since $\hat{p}$ is approximately $N(p_0, p_0(1-p_0)/n)$, Z will be approximately standard normal, This means

$$P\left(\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} > z_\alpha \middle| \mu=\mu_0\right) \approx \alpha \tag{3.4a}$$

$$P\left(\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} < -z_\alpha \middle| \mu=\mu_0\right) \approx \alpha \tag{3.4b}$$

and

$$P\left(\left[\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}<-z_{\alpha/2}\right]\bigcup\left[\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}>z_{\alpha/2}\right]\middle|\mu=\mu_0\right)\approx\alpha \quad (3.4c)$$

That makes the large-sample tests

1. If H_a is $p>p_0$, we reject H_0 if $\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}>z_\alpha$.
2. If H_a is $p<p_0$, we reject H_0 if $\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}<-z_\alpha$.
3. If H_a is $p\neq p_0$, we reject H_0 if either $\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}<-z_{\alpha/2}$ or $\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}>z_{\alpha/2}$.

3.1.1 False Dismissal Probability

Estimating $\beta(p')$ for these tests as a function of the actual proportion p' is a little different than in the case of a population mean, since now the variance depends on the parameter p' as well, i.e., if $p=p'$, we know $E(\hat{p})=p'$ and $V(\hat{p})=p'(1-p')/n$, so

$$E(Z)=\frac{p'-p_0}{\sqrt{p_0(1-p_0)/n}} \quad (3.5)$$

and

$$V(Z)=\frac{p'(1-p')/n}{p_0(1-p_0)/n}=\frac{p'(1-p')}{p_0(1-p_0)} \quad (3.6)$$

So for example if H_a is $p>p_0$, we have false dismissal probability

$$\begin{aligned}\beta(p')=P\left(Z\le z_\alpha\middle|\,p=p'\right)&=\Phi\left(\frac{z_\alpha-(p'-p_0)/\sqrt{p_0(1-p_0)/n}}{\sqrt{[p'(1-p')]/[p_0(1-p_0)]}}\right)\\&=\Phi\left(\frac{z_\alpha\sqrt{p_0(1-p_0)/n}-(p'-p_0)}{\sqrt{p'(1-p')/n}}\right)\end{aligned} \quad (3.7)$$

3.2 Small Sample Tests

If the sample size is too small (or p_0 is to close to zero or one) to use the normal trick, we basically have to construct the test using the binomial cdf

$$B(x;n,p)=\sum_{y=0}^{x}b(x;n,p)=\sum_{y=0}^{x}\binom{n}{x}p^x(1-p)^{n-x} \quad (3.8)$$

In practice we won't actually evaluate the sum; we'll look it up in a table or ask a statistical software package to do it for us.

A test which rejects H_0 when $X\ge c$, i.e., $\hat{p}\ge c/n$, appropriate for alternative hypothesis H_a: $p>p_0$, will have a false alarm probability of

$$\alpha=P(X\!\ge\!c|p\!=\!p_0)=1-P(X\!\le\!c-1|p\!=\!p_0)=1-B(c-1;n,p_0) \quad (3.9)$$

and similarly for lower-tailed and two-tailed test. In general, we won't be able to produce a test with exactly the desired false alarm probability, but we can pick one which is close.

Practice Problems

8.35, 8.39, 8.43, 8.45

Thursday 22 September 2016

Review for Prelim Exam One (up to and including section 8.2). Please bring questions, and ideally ask them by email before class.

Tuesday 27 September 2016

Prelim Exam One (up to and including section 8.2). Closed book, closed notes, but you may bring one handwritten

8.5"×11" (front and back) formula sheet, and also use a scientific calculator.

Thursday 29 September 2016

4 P-values

So far we've carried out hypothesis testing by specifying the desired false alarm probability (significance) α for the test, and then based on the statistic calculated from the data, getting a yes or no answer on rejecting the null hypothesis. But this doesn't capture, for example, whether we just managed to reject H_0 at that level, or cleared the threshold by a good margin. For example, suppose we construct a z statistic for a one-tailed test, and obtain the value $z = 1.84$. If our test was designed to have false-alarm probability $\alpha = 0.10$, the threshold would be $z_{.10} = 1.282$, and since $1.84 > 1.282$, we would reject H_0 according to this test. On the other hand, if we'd chosen $\alpha = 0.01$, the threshold would be $z_{.01} = 2.326$, and since $1.84 < 2.326$, we would not reject H_0 using the test at this lower false alarm probability. Looking at some levels in between, we find

α	.10	.05	.025	.01
z_α	1.282	1.645	1.960	2.326
Reject H_0?	Yes	Yes	No	No

If the false alarm rate for the test is "low" given the data, we do not reject the null hypothesis; if it is "high", we do. Evidently, there's some dividing value of α at which $z = z_\alpha$; at this false alarm probability we cross over from rejecting to not rejecting H_0. If $z_\alpha = 1.84$, this defines the α in question as

$$\alpha = P(Z > z_\alpha) = 1 - \Phi(z_\alpha) = 1 - \Phi(1.84) \approx 1 - .9671 = .0329 \tag{4.1}$$

we call this critical α the P-value for the hypothesis comparison, given the data. In general, we can say

The P-value associated with a data set in the context of a family of hypothesis tests is the highest false alarm probability at which we fail to reject the null hypothesis.

Another way of saying it is, it's the probability, if the null hypothesis is true, that you'd obtain data at least as inconsistent with the null hypothesis as what you actually observed. For a lower-tailed test, this means it's the probability of obtaining a statistic lower than the actual value; for instance with a lower-tailed t test with 5 degrees of freedom and a statistic value of -2.7, the P value is (the value can be obtained from table A-8 of Devore)

$$P(T_5 < -2.7) = F_T(-2.7, 6) = .018 \tag{4.2}$$

For a two-tailed Z or T test, it's the probability that the statistic would be as far away from zero as what you observed. For example, if we found $z = 1.84$ in the context of a two-sided alternative hypothesis, we'd have a P value of

$$P(Z < -1.84) + P(Z > 1.84) = .0329 + .0329 = .0658 \tag{4.3}$$

Similarly, for a t statistic of -2.7, where there are five dof and a two-sided alternative hypothesis, we have a P value of

$$P(T_5 < -2.7) + P(T_5 > 2.7) = .018 + .018 = .036 \tag{4.4}$$

Explicitly, for a z test, the P value is

1. $1 - \Phi(z)$ for an upper-tailed test
2. $\Phi(-z)$ for a lower-tailed test
3. $\Phi(-|z|) + 1 - \Phi(|z|) = 2\Phi(-|z|)$ for a two-tailed test

5 Bonus Material

5.1 ROC Curves

5.2 Likelihood Ratio Tests

Practice Problems

8.49, 8.51, 8.81, 8.87