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

MATH-252-01: Probability and Statistics II*

Spring 2019

Contents

Tuesday 5 February 2019

Note that the structure of Chapter 8 has changed somewhat between the eighth and ninth editions of Devore, due mostly to a pedagogical choice to make P-values more central to the discussion. As this is a reasonable choice, we'll be making the same shift in lecture. For an alternative treatment which introduces P-values later, see the previous notes at `http://ccrg.rit.edu/~whelan/courses/2016_3fa_MATH_252/notes08.pdf`

1 Hypothesis Tests (illustrated with z-tests)

1.1 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.

1.2 Example: z tests for population mean

For example, consider a sample of size n drawn from a normal distribution with unknown mean μ and known variance σ^2. Previously, we used the fact that the sample mean $\overline{X}$ is a $N(\mu, \sigma^2)$ random variable to define a pivot variable $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ which was known to be standard normal. Now, we'll evaluate a null hypothesis H_0 which says that $\mu = \mu_0$ where μ_0 is some specified value (which may be zero, but need not be). There are three choices of alternative hypothesis H_a: $\mu \neq \mu_0$, $\mu > \mu_0$, and $\mu < \mu_0$. In each case the test statistic will be

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \tag{1.1}$$

If H_0 is true, this is again standard normal [$N(0,1)$]. If H_a is true, Z will still be a normal random variable with a variance of one, but its mean will be $\frac{\mu-\mu_0}{\sigma/\sqrt{n}}$. Depending on the choice of alternative hypothesis, $E(Z)$ might be known to be positive, known to be negative, or simply non-zero.

The way we conduct the test is to convert the actual data $\{x_i\}$ into an actual value $z = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}$ where $\overline{x}$ is the actual observed sample mean. Since we know that this test statistic Z is standard normal if the null hypothesis H_0 is true, we should begin to doubt H_0 if the observed z is too far from zero. Depending on the form of the alternative hypothesis H_a, we could reject H_0 if z were too high, or too low, or perhaps either.

1.3 P-value

The decision to reject the null hypothesis H_0 or not is based on whether the data seem consistent with H_0. We quantify that with something called the P value. This is the probability that data generated according to the null hypothesis would generate a test statistic value at least as "extreme" as the actual value observed.

1.3.1 Case when H_a: $\mu > \mu_0$

For concreteness, suppose the alternative hypothesis H_a is that $\mu > \mu_0$. Then H_a tells us the test statistic Z is a normal random variable with a variance of one and a *positive* mean. Thus a sufficiently high (positive) value of z will look inconsistent with the null hypothesis H_0 and consistent with the alternative hypothesis H_a. (A very negative value of z will look consistent with neither hypothesis, but more like H_0 than H_a.) We define the P value as the chance of finding a test statistic value of z or higher if H_0 is true:

$$P = P(Z \geq z | H_0 \text{ is true}) = 1 - \Phi(z) = \Phi(-z) \tag{1.2}$$

This is the upper tail probability (Table A.3 of Devore). For instance, if $z = -1$, $P \approx 0.8413$, but if $z = 3$, $P \approx 0.0013$. The lower the P value, the more inconsistent the data are with H_0.

1.3.2 Case when H_a: $\mu < \mu_0$

If the alternative hypothesis is $\mu < \mu_0$, then it's low values of Z which are considered anomolous (in a way consistent with the alternative hypothesis), so the P value corresponding to a z score is the probability, under the null hypothesis, of finding a lower (or more negative) value:

$$P = P(Z \leq z | H_0 \text{ is true}) = \Phi(z) \tag{1.3}$$

1.3.3 Case when H_a: $\mu \neq \mu_0$

Now we can consider either high or low z scores to be anomolous, so the P value is the probability of finding Z farther from zero than z, assuming the null hypothesis. So if $z = 2.5$, this is the probability that $Z \geq 2.5$ or $Z \leq -2.5$. If $z = -1.2$, it is the probability that $Z \leq -1.2$ or $Z \geq 1.2$. This can be summarized as

$$\begin{aligned} P &= P(|Z| \geq |z| \, | H_0 \text{ is true}) \\ &= P(Z \leq -|z| \, | H_0 \text{ is true}) + P(Z \geq |z| \, | H_0 \text{ is true}) \\ &= \Phi(-|z|) + (1 - \Phi(|z|)) = \Phi(-|z|) + \Phi(-|z|) \\ &= 2\Phi(-|z|) \end{aligned} \tag{1.4}$$

1.4 Type I and Type II Errors

To get back to the concept of a test as a way to decide between rejecting H_0 and not rejecting it, this statistic-based method actually defines a whole family of tests. For any value of α between 0 and 1, we can define a test which rejects H_0 if $P \leq \alpha$. (Recall that lower P values are "worse news" for H_0.) The value α has many names, but a common one is "significance". Note that this name can be a bit confusing, because a test with a significance of $\alpha = .01$ will be more stringent (require more evidence to reject H_0 than one with a significance of $\alpha = .05$.

Because of the random nature of the experiment, there will be some probability that a given 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

$$P(\text{reject } H_0|H_0 \text{ is true}) = P(P \leq \alpha|H_0 \text{ is true}) \tag{1.5}$$

Since the definition of the P value implies that, if H_0 is true, it will be a uniform random variable between 0 and 1 (e.g., we should find $P \leq .2$ twenty percent of the time), the false alarm probability is actually just the significance α:

$$P(\text{reject } H_0|H_0 \text{ is true}) = P(P \leq \alpha|H_0 \text{ is true}) = \alpha \tag{1.6}$$

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

$$\beta = 1 - P(\text{reject } H_0|H_a \text{ is true}) \tag{1.7}$$

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

$$\beta(\theta) = 1 - P(\text{reject } H_0|\text{parameter value } \theta) \tag{1.8}$$

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 $1 - \beta(\theta)$.

1.5 Specifics of z test

Returning to the case where we have a sample of size n drawn from a normal distribution (or population) with unknown mean μ and known variance σ^2, and null hypothesis H_0: $\mu = \mu_0$, we consider the threshold on $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ corresponding to a desired false alarm probability α. As noted, this corresponds in each case to a test which rejects H_0 if $P \leq \alpha$. Since

$$P = 1 - \Phi(z) = P(Z \geq z|H_0 \text{ is true}) \tag{1.9}$$

and $1 - \Phi(z_\alpha) = \alpha$ by definition, the threshold of α on P is a threshold of z_α on z. Similar calculations in the other two cases tell us:

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.

Practice Problems

8.1, 8.7, 8.9, 8.13, 8.19

Thursday 7 February 2019

1.6 False Dismissal Probability

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 $\mu > \mu_0$

$$\beta(\mu') = P\left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \leq z_\alpha \middle| \mu = \mu'\right) = \Phi\left(z_\alpha - \frac{\mu' - \mu_0}{\sigma/\sqrt{n}}\right) \tag{1.10}$$

while if H_a is $\mu < \mu_0$

$$\begin{aligned} \beta(\mu') &= P\left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \geq -z_\alpha \middle| \mu = \mu'\right) = 1 - \Phi\left(-z_\alpha - \frac{\mu' - \mu_0}{\sigma/\sqrt{n}}\right) \\ &= \Phi\left(z_\alpha - \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) \end{aligned} \tag{1.11}$$

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:

$$\begin{aligned} \beta(\mu') &= P\left(-z_{\alpha/2} \leq \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \middle| \mu = \mu'\right) \\ &= \Phi\left(z_{\alpha/2} - \frac{\mu' - \mu_0}{\sigma/\sqrt{n}}\right) - \Phi\left(-z_{\alpha/2} - \frac{\mu' - \mu_0}{\sigma/\sqrt{n}}\right) \end{aligned} \tag{1.12}$$

1.7 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

$$\beta = 1 - \Phi(z_\beta) = \Phi(-z_\beta)\ , \tag{1.13}$$

which means that

$$-z_\beta = \begin{cases} z_\alpha - \frac{\mu' - \mu_0}{\sigma/\sqrt{n}} & \text{upper tailed} \\ z_\alpha - \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} & \text{lower tailed} \end{cases} \tag{1.14}$$

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

$$n = \left(\frac{\sigma(z_\alpha + z_\beta)}{\mu' - \mu_0}\right)^2 \tag{1.15}$$

2 t-tests (unknown variance)

Now suppose we have a sample of size n drawn from a normal distribution, but both the mean and the variance is unknown. If we want to assess the null hypothesis H_0: $\mu = \mu_0$, we can't construct the usual test statistic $\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$ because the standard deviation σ is unknown. As in the case of confidence intervals, we use the sample variance as an estimate. The resulting test statistic

$$T = \frac{\overline{X} - \mu_0}{\sqrt{S^2/n}} \tag{2.1}$$

is Student-t-distributed with $n - 1$ degrees of freedom. (Again, this is a consequence of Student's Theorem.) We then reject the null hypothesis if the statistic T is too far from zero in the appropriate direction. Specifically, if we want to define tests at a fixed significance α, they go as follows:

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

To describe the outcome in terms of the P-value, we need to use the cumulative distribution function of the t-distribution,

$F_T(t; n-1)$ or equivalently the tail probability (area to the right of t $1 - F_T(t; n-1) = F_T(-t; n-1)$. Then, defining $t = \frac{\overline{x}-\mu_0}{s/\sqrt{n}}$,

1. If H_a is $\mu > \mu_0$, $P = F_T(-t; n-1) = \int_t^\infty f_T(u; n-1)\, du$
2. If H_a is $\mu < \mu_0$, $P = F_T(t; n-1) = \int_{-\infty}^t f_T(u; n-1)\, du$
3. If H_a is $\mu \neq \mu_0$, $P = 2F_T(-|t|; n-1) = \int_{|t|}^\infty f_T(u; n-1)\, du$

Devore has the t tail probability

$$\int_t^\infty f_T(u; \nu)\, du \tag{2.2}$$

for various numbers of degrees of freedom tabulated.

Note that the false dismissal probability $\beta(\mu)$ (or the power $\gamma(\mu) = 1 - \beta(\mu)$) is not so easy to calculate for a t-test. This is because $T = \frac{\overline{X}-\mu_0}{\sqrt{S^2/n}}$ doesn't have a simple probability distribution when $\mu \neq \mu_0$.

2.1 Large-Sample Approximation

As with confidence intervals, we can take advantage of the fact that, as the number of degrees of freedom becomes large, the Student-t distribution approaches the standard normal distribution. So if $n > 40$ or so, we can assume the test statistic $\frac{\overline{X}-\mu_0}{\sqrt{S^2/n}}$ is approximately standard normal (in fact, it's common to call it Z rather than T in this case) and

1. If H_a is $\mu > \mu_0$, then $P \approx 1 - \Phi(\frac{\overline{x}-\mu_0}{s/\sqrt{n}})$ and we reject H_0 if $\frac{\overline{X}-\mu_0}{\sqrt{S^2/n}} > z_\alpha$.
2. If H_a is $\mu < \mu_0$, then $P \approx \Phi(\frac{\overline{x}-\mu_0}{s/\sqrt{n}})$ and we reject H_0 if $\frac{\overline{X}-\mu_0}{\sqrt{S^2/n}} < -z_\alpha$.
3. If H_a is $\mu \neq \mu_0$, then $P \approx 2\Phi(-|\frac{\overline{x}-\mu_0}{s/\sqrt{n}}|)$ and we reject H_0 if either $\frac{\overline{X}-\mu_0}{\sqrt{S^2/n}} < -z_{\alpha/2}$ or $\frac{\overline{X}-\mu_0}{\sqrt{S^2/n}} > z_{\alpha/2}$.

Thanks to the central limit theorem, we don't even need to know that the underlying distribution is normal in the large-sample case. As long as it has a finite variance, the test statistic based on $\overline{X}$ and S^2 will be approximately standard normal.

3 Hypothesis Tests and Confidence Intervals

You may have noticed that the procedures involved in conducting a hypothesis test at specified significance α and constructing a confidence interval of specified confidence level $1 - \alpha$ are similar. Let's examine that more closely, using the example of a sample of size n drawn from a normal distribution with unknown μ and σ. The two-sided confidence interval will then be

$$\text{from} \quad \overline{x} - t_{\alpha/2,n-1}\sqrt{s^2/n} \quad \text{to} \quad \overline{x} + t_{\alpha/2,n-1}\sqrt{s^2/n} \tag{3.1}$$

On the other, hand the rules of a two-tailed hypothesis test say

$$\text{Reject } H_0 \quad \text{if} \quad \frac{\overline{x}-\mu_0}{\sqrt{s^2/n}} \geq t_{\alpha/2,n-1} \quad \text{or} \quad \frac{\overline{x}-\mu_0}{\sqrt{s^2/n}} \leq -t_{\alpha/2,n-1} \tag{3.2}$$

or equivalently

$$\text{Reject } H_0 \quad \textit{unless} \quad -t_{\alpha/2,n-1} < \frac{\overline{x}-\mu_0}{\sqrt{s^2/n}} < t_{\alpha/2,n-1} \tag{3.3}$$

A little bit of algebra shows this is equivalent to

$$\text{Reject } H_0 \quad \textit{unless} \quad \overline{x}-t_{\alpha/2,n-1}\sqrt{s^2/n} < \mu_0 < \overline{x}+t_{\alpha/2,n-1}\sqrt{s^2/n} \tag{3.4}$$

I.e., you reject the null hypothesis if the specified value μ_0 lies *outside* the corresponding confidence interval.

Practice Problems

8.31, 8.37, 8.39, 8.55

Tuesday 12 February 2019

4 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 $\mathrm{Bin}(1,p)$, or a single binomial random variable $X \sim \mathrm{Bin}(n,p)$. In any event, it's more convenient to work with the estimator $\hat{p} = X/n$, which has mean

$$E(\hat{p}) = \frac{np}{n} = p \tag{4.1}$$

and variance

$$V(\hat{p}) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n} \tag{4.2}$$

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.

4.1 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

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \tag{4.3}$$

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

$$P\left(\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} > z_\alpha \middle| \mu = \mu_0\right) \approx \alpha \tag{4.4a}$$

$$P\left(\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} < -z_\alpha \middle| \mu = \mu_0\right) \approx \alpha \tag{4.4b}$$

and

$$P\left(\left[\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} < -z_{\alpha/2}\right] \bigcup \left[\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} > z_{\alpha/2}\right] \middle| \mu = \mu_0\right) \approx \alpha \tag{4.4c}$$

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 $\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} > z_{\alpha/2}$.

4.1.1 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

$$E(Z) = \frac{p' - p_0}{\sqrt{p_0(1-p_0)/n}} \tag{4.5}$$

and

$$V(Z) = \frac{p'(1-p')/n}{p_0(1-p_0)/n} = \frac{p'(1-p')}{p_0(1-p_0)} \tag{4.6}$$

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

$$\begin{aligned}\beta(p') &= P\left(Z \le z_\alpha \middle| p = p'\right) = \Phi\left(\frac{z_\alpha - (p'-p_0)/\sqrt{p_0(1-p_0)/n}}{\sqrt{[p'(1-p')]/[p_0(1-p_0)]}}\right) \\ &= \Phi\left(\frac{z_\alpha\sqrt{p_0(1-p_0)/n} - (p'-p_0)}{\sqrt{p'(1-p')/n}}\right)\end{aligned} \tag{4.7}$$

4.2 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

$$B(x; n, p) = \sum_{y=0}^{x} b(x; n, p) = \sum_{y=0}^{x} \binom{n}{x} p^x (1-p)^{n-x} \tag{4.8}$$

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

$$\alpha = P(X \ge c | p = p_0) = 1 - P(X \le c-1 | p = p_0) = 1 - B(c-1; n, p_0) \tag{4.9}$$

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