The Variability of P-Values

Dennis D. Boos
Department of Statistics
North Carolina State University
Raleigh, NC 27695-8203
boos@stat.ncsu.edu

August 15, 2009

NC State Statistics Departement Tech Report # 2626

Summary

P-values are used as evidence against a null hypothesis. But is it enough to report a p-value without some indication of its variability? Bootstrap prediction intervals and standard errors on the log scale are proposed to complement the standard p-value.

Key words and phrases: prediction interval, log p-value, measure of evidence.

1 Introduction

Good statistical practice is to report some measure of variability or reliability for important quantities estimated, that is, for estimates of the primary parameters of interest. Thus if the population mean μ is a key focus and estimated by the sample mean $\overline{Y}$ from a simple iid sample $Y_1, \ldots, Y_n$, one should also report the standard error $s/\sqrt{n}$ and/or a confidence interval for μ. One could carry the logic forward and say that standard errors should also be reported for $s/\sqrt{n}$ or for the endpoints of the confidence interval, but most would agree that standard errors and confidence interval endpoints are secondary measures that usually do not require standard errors themselves.

A p-value is simply the probability of obtaining as extreme or a more extreme result than found in observed data $\boldsymbol{Y}$, where the probability is computed under a null hypothesis H_0. Thus, for example, if T is a statistic for which large values suggest rejection of H_0, then the p-value is just $P(T \geq T(\boldsymbol{Y}) \mid \boldsymbol{Y})$, where $T(\boldsymbol{Y})$ is the value of T computed from the observed data, considered fixed, and T is an independent generation of T under H_0.

Are p-values primary measures and deserving of some focus on their variability? Perhaps the answer has to do with how much they are emphasized and relied upon, but often p-values are thought of as important measures of evidence relative to H_0. In that role, it seems plausible to understand something of their variability.

I use p-values as a quick and dirty method for detecting statistical significance, i.e., true "signals" in the presence of noise. In recent years, I have been promoting the use of exact p-values because there seems to be no reason to rely on mathematical approximation when exact methods are available. For example, rank statistics and categorical data analyses are often built on permutation methods that allow conditionally exact p-values. Statistical computing packages like StatXact have led the way in making these p-values available, at least by Monte Carlo approximation.

However, I recently realized that promoting the use of exact p-values is perhaps questionable without knowing more about the variability of p-values. That caused me to think more

deeply about the problem. I get excited when I see a p-value like 0.002. But what if an identical replication of the experiment could lead to p-values that range from 0.001 to 0.15? That is not so exciting.

So, how should one measure and report the variability of a p-value? This is not as obvious as it first seems. For continuous T under a simple H_0, the p-value has a uniform(0,1) distribution. But under an alternative H_a, the logarithm of the p-value is asymptotically normal (see Lambert and Hall, 1982, and Example 2 below). Thus, the p-value can have a quite skewed distribution under H_a, and simply reporting an estimate of its standard deviation is not very satisfying for those situations. One solution is to get comfortable with the log(p-value) scale and just report a bootstrap or jackknife standard error on that scale. (Actually, I prefer the base 10 logarithm as I will explain later.)

But because most people are used to thinking on the p-value scale and not its logarithm, I propose a prediction interval for the p-value composed of modified quantiles from the bootstrap distribution of the p-value. This bootstrap prediction interval captures the p-value from a new, independent replication of the experiment. Mojirsheibani and Tibshirani (1996) and Mojirsheibani (1998) developed these bootstrap intervals for a general statistic computed from a future sample. One of their intervals, the nonparametric bias-corrected (BC) bootstrap prediction interval, is briefly described in Section 3 and used in the examples in Section 2 and in the Monte Carlo experiments in Section 4.

2 Examples

Example 1. To decide if consumers will be able to distinguish between two types of a food product, $n = 50$ sets of three samples are prepared, with half of the sets consisting of two samples from type A and one from type B, and half of the sets with two samples from Type B and one from Type A. Fifty randomly chosen judges are asked to tell which samples of their set are from the same type. The hypotheses are $H_0 : \pi = 1/3$ (random guessing) versus $H_a : \pi > 1/3$. Suppose that 25 judges are successful in identifying the pairs with the same type. The exact p-value from the binomial distribution is $P(\text{Bin}(n = 50, p = 1/3) \geq 25) =$

0.011, and the normal approximation without continuity correction is 0.006. Should we care that the normal approximation overstates the significance by roughly a factor of 2? Probably not. Let's see why.

Table 1 gives p-value prediction intervals when $\widehat{\pi} = 1/2$ and $n = 20$, 50, and 100 using the bootstrap prediction interval method described in Section 3. For example, an 80% prediction interval when $n = 50$ is (0.000031, 0.29) Thus, a replication of the experiment will result in a p-value lying in this interval 80% of the time. The related 90% lower and upper bounds are 0.000031 and 0.29, respectively. The central 50% interval is (0.00082, 0.076), encompassing values much farther from the exact p-value = 0.011 than the normal approximation p-value = 0.006.

Table 1: Bootstrap prediction interval endpoints for a future p-value and standard errors for the base 10 logarithm of the p-value. Binomial test of $H_0 : \pi = 1/3$ versus $H_a : \pi > 1/3$, $\widehat{\pi} = 1/2$.

n	20	50	100
p-value	0.092	0.011	0.00042
level	Interval Endpoints		
0.10	0.00088	0.000031	0.00000013
0.25	0.013	0.00082	0.0000069
0.75	0.70	0.076	0.010
0.90	0.85	0.29	0.066
$-\log_{10}(p\text{-value})$	1.04	1.97	3.38
bootstrap se	0.85	1.20	1.61
jackknife se	0.77	1.15	1.57
95% c.i. for π	(0.29,0.71)	(0.36,0.64)	(0.40,0.60)

Bootstrap endpoints and standard errors computed from binomial(n,$\widehat{\pi}$) without resampling. Endpoints are one-sided bounds. Intervals are formed by combining symmetric endpoints (0.10,0.90) or (0.25,0.75).

Table 1 also reports the negative of the base 10 logarithm of the p-value, $-\log_{10}(p\text{-value})$, and the bootstrap and jackknife standard errors on this scale. (Note that for positive x, $\log_{10}(x) = 0.4343 \log_e(x)$.) Why use base 10? Recall the star (*) method of reporting

significance results used in many subject matter journals: one star * for p-values ≤ 0.05, two stars ** for p-values ≤ 0.01, and three stars *** for p-values ≤ 0.001. If we replace 0.05 by 0.10, then the rule is: * if $-\log_{10}(p\text{-value})$ is in the range [1,2), ** for the range [2,3), and *** for the range [3,4). In fact, looking at the standard errors for $-\log_{10}(p\text{-value})$ in Table 1, it might make sense to round the $-\log_{10}(p\text{-value})$ values to integers. That would translate into rounding the p-values: 0.092 to 0.10, 0.011 to 0.01, and 0.00042 to 0.001. Thus, taking into account the variability in $-\log_{10}(p\text{-value})$ seems to lead back to the star (*) method. To many statisticians this rounding seems information-robbing, but this opinion may result from being unaware of the variability inherent in p-values.

A standard complement to the p-value is a 95% confidence interval for π. With $\widehat{\pi} = 1/2$, these intervals for $n = 20$, 50, and 100 are given at the bottom of Table 1. They certainly help counter over-emphasis on small p-values. For example, at $n = 50$ the confidence interval is (0.36, 0.64), showing that the true π could be quite near the null value 1/3.

Table 2 illustrates the prediction intervals when the p-value is held constant near 0.0038, but n changes from 20 to 54 and to 116. These results suggest that the same p-value at different sample sizes has a somewhat similar variability and repeatability. For example, with a p-value around 0.0038 for these situations, the 75% upper bounds are roughly similar: 0.038, 0.058, and 0.063. Note also that the bootstrap and jackknife standard errors in Table 2 for the base 10 logarithm of the p-value are nearly the same. The confidence intervals for π, though, are quite different in these experiments because they are centered differently.

Example 2. Consider samples of size n from a normal distribution with mean μ and variance $\sigma^2 = 1$ and the hypotheses $H_0 : \mu = 0$ vs. $H_a : \mu > 0$. If one assumes that the variance is known, then Lambert and Hall (1982, Table 1) give that the p-values from $Z = \sqrt{n}\overline{Y}/\sigma$ for $\mu > 0$ in H_a satisfy

$$\sqrt{n}\left(\frac{\log(p\text{-value})}{n} + \mu^2/2\right) \xrightarrow{d} \mathrm{N}(0, \mu^2), \qquad \text{as } n \to \infty, \tag{1}$$

or equivalently that $-\log(p\text{-value})/n$ is asymptotically normal with asymptotic mean $\mu^2/2$ and asymptotic variance μ^2/n. This asymptotic normality implies that

$$-\log(p\text{-value})/n \xrightarrow{p} \mu^2/2, \qquad \text{as } n \to \infty. \tag{2}$$

Table 2: Bootstrap prediction interval endpoints for a future p-value and standard errors for the base 10 logarithm of the p-value. Binomial test of $H_0 : \pi = 1/3$ versus $H_a : \pi > 1/3$.

n	20	54	116
Y	13	28	53
p-value	0.0037	0.0038	0.0038
level	Interval Endpoints		
0.10	0.0000028	0.0000022	0.0000027
0.25	0.00017	0.00024	0.00015
0.75	0.038	0.058	0.043
0.90	0.19	0.23	0.22
$-\log_{10}(p\text{-value})$	2.43	2.42	2.42
bootstrap se	1.28	1.34	1.35
jackknife se	1.24	1.29	1.30
95% c.i. for π	(0.41,0.85)	(0.38,0.66)	(0.36,0.55)

Bootstrap endpoints and standard errors computed from binomial(n,$\widehat{\pi}$) without resampling. Endpoints are one-sided bounds. Intervals are formed by combining symmetric endpoints (0.10,0.90) or (0.25,0.75).

A similar result obtains in the unknown variance case where the usual $t = \sqrt{n}\overline{Y}/s$ replaces Z, but the constants are more complicated. For example, the asymptotic mean is $(1/2)\log(1 + \mu^2/2)$, in place of $\mu^2/2$, and the asymptotic variance is $\mu^2(1+\mu^2)^{-2}(1+\mu^2/2)$ in place of μ^2.

Table 3 illustrates the asymptotic normality of $-\log_{10}(p\text{-value})$ for the one-sample t-test. The p-values themselves are clearly not normally distributed because the third moment ratio Skew and the fourth moment ratio Kurt given by $\mathrm{E}\{X - \mathrm{E}(X)\}^k / \left[\mathrm{E}\{X - \mathrm{E}(X)\}^2\right]^{k/2}$ for $k = 3$ and $k = 4$, are not near their normal distribution values of 0 and 3, respectively. Although the Skew values for $-\log_{10}(p\text{-value})$ are not close to 0, the trend from $n = 10$ to $n = 20$ is down, and the values at $n = 50$ (not displayed) are 0.60 and 0.33, respectively, for $\mu = 0.5$ and $\mu = 1.0$.

Can we obtain an approximate prediction interval from the asymptotic normality? In-

Table 3: Distribution summaries for the t-test p-value and $-\log_{10}(p\text{-value})$ for the normal location problem, $H_0 : \mu = 0$ versus $H_a : \mu > 0$, $\sigma = 1$ unknown.

	p-value				$-\log_{10}(p\text{-value})$			
	Mean	SD	Skew	Kurt	Mean	SD	Skew	kurt
$n = 10, \mu = 0.5$	0.14	0.17	1.87	6.5	1.29	0.77	0.99	4.2
$n = 10, \mu = 1.0$	0.020	0.044	5.6	49	2.39	0.91	0.58	3.5
$n = 20, \mu = 0.5$	0.062	0.109	3.25	16.1	1.89	0.98	0.85	4.1
$n = 20, \mu = 1.0$	0.002	0.008	14	280	4.07	1.26	0.47	3.4

Based on 10,000 Monte Carlo samples. Skew and Kurt are the moment ratios, $\mathrm{E}\{X - \mathrm{E}(X)\}^k / \left[\mathrm{E}\{X - \mathrm{E}(X)\}^2\right]^{k/2}$, for $k = 3$ and $k = 4$, respectively. Standard errors of entries are in the last decimal displayed or smaller.

verting the asymptotic pivotal statistic implied by (1) leads to the asymptotic prediction interval

$$\left(\exp\left\{-z_{1-\alpha/2}\sqrt{n}\mu - n\mu^2/2\right\},\ \exp\left\{z_{1-\alpha/2}\sqrt{n}\mu - n\mu^2/2\right\}\right), \tag{3}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal. Substituting the sample mean $\overline{Y}$ for μ then yields a possible competitor to the bootstrap intervals. Unfortunately, unless $\overline{Y}$ is fairly large, the right endpoint is often larger than one, and the interval is not informative.

However, the result (1) seems to support the observation from Table 2 that equal p-values from different experiments with different sample sizes have similar prediction intervals. That is, letting P_n=p-value and setting $\mu^2/2 = -\log(P_n)/n$ from (2) and solving for μ yields $\widehat{\mu} = \{-2\log(P_n)/n\}^{1/2}$. Then, plugging $\widehat{\mu}$ into (3) yields

$$\left(\exp\left\{-z_{1-\alpha/2}\left\{-2\log(P_n)\right\}^{1/2} + \log(P_n)\right\},\ \exp\left\{z_{1-\alpha/2}\left\{-2\log(P_n)\right\}^{1/2} + \log(P_n)\right\}\right), \tag{4}$$

and sample sizes do not appear in the endpoint expressions except through P_n.

The general result in Lambert and Hall (1982) for a variety of testing procedures about μ is that $-\log(p\text{-value})$ is asymptotically normal with asymptotic mean $-nc(\mu)$ and asymp-

totic variance $n\tau^2(\mu)$, where the functions $c(\cdot)$ and $\tau^2(\cdot)$ depend on the procedure. Thus, the asymptotic mean is proportional to n, and the asymptotic standard deviation is proportional to $\sqrt{n}$. In Table 3, these results are crudely supported by the ratio of means for $-\log_{10}(p\text{-value})$, 1.89/1.29=1.5 for $\mu = 0.5$ and 4.07/2.39=1.7 for $\mu = 1.0$ where $20/10 = 2$. Similarly, the ratios of standard deviations 0.98/0.77=1.3 and 1.26/0.91=1.4 are not far from $\sqrt{20/10} = 1.4$.

Example 3. Miller (1986, p. 65) gives data from a study comparing partial thromboplastin times for patients whose blood clots were dissolved (R=recanalized) and for those whose clots were not dissolved (NR):

R	41	86	90	74	146	57	62	78	55	105
	46	94	26	101	72	119	88			
NR	34	23	36	25	35	23	87	48		

The exact two-sided Wilcoxon Rank Sum p-value is 0.0014 and the normal approximation without continuity correction is 0.0024. A 50% prediction interval for an exact future p-value based on $B = 999$ resamples is (0.00007, 0.018) and for the approximate p-value is (0.00037, 0.020). The value of $-\log_{10}(p\text{-value})$ for the exact p-value is 2.84 with bootstrap standard error 1.27 and jackknife standard error 1.31, thus suggesting one round 2.84 to 3 and the p-value from 0.0014 to 0.001. Similarly, the value of $-\log_{10}(p\text{-value})$ for the normal approximation p-value is 2.61 with bootstrap standard error 0.84 and jackknife standard error 0.89, leading to rounding 2.61 to 3 and 0.0024 to 0.001.

Example 4. The following data, found in Larsen and Marx (2001, p. 15), are the ordered values of a sample of width/length ratios of beaded rectangles of Shoshoni indians:

.507	.553	.576	.601	.606	.606	.609	.611	.615	.628
.654	.662	.668	.670	.672	.690	.693	.749	.844	.933

Suppose that we wanted to use the Anderson-Darling goodness-of-fit statistic to test whether the data are normally distributed. Using the accurate approximation for the p-value taken from Stephens (1986, Table 4.9), the p-value is 0.012. A 90% nonparametric bootstrap prediction interval for a new p-value is $(0.000027, 0.36)$, and the 50% interval is $(0.00055, 0.094)$,

based on $B = 999$ resamples. So the p-value from a second sample from this population could easily be higher than the observed 0.012, but we expect it to be less than 0.094 *with confidence* 0.75. The value of $-\log_{10}(p\text{-value})$ is 1.93 with bootstrap standard error 1.23, again suggesting one round 1.93 to 2 and the p-value from 0.012 to 0.01.

3 Bootstrap Prediction Intervals

Consider an iid random sample $Y_1, \ldots, Y_n$ and an independent "future" iid sample $X_1, \ldots, X_m$ from the same population, and statistics T_n and T_m computed from these samples, respectively. If $n = m$, then T_n and T_m have identical distributions. Mojirsheibani and Tibshirani (1996) and Mojirsheibani (1998) derived bootstrap prediction intervals from the Y sample that contain T_m with approximate probability $1 - \alpha$. Here I briefly describe their bias-corrected (BC) interval. Mojirsheibani and Tibshirani (1996) actually focused on bias-corrected accelerated (BC_a) intervals, but I am using the BC intervals for simplicity.

Let $Y_1^*, \ldots, Y_n^*$ be a random resample taken with replacement from the set $(Y_1, \ldots, Y_n)$, i.e., a nonparametric bootstrap resample, and let $T_n^{(1)}$ be the statistic calculated from the resample. Repeating the process independently B times results in $T_n^{(1)}, \ldots, T_n^{(B)}$. Let $\widehat{K}_B$ be the empirical distribution function of these $T_n^{(i)}$, and let $\widehat{\eta}_B(\alpha)$ be the related αth sample quantile. Then, the $1 - \alpha$ bias-corrected (BC) bootstrap percentile prediction interval for T_m is

$$\{\widehat{\eta}_B(\alpha_1),\ \widehat{\eta}_B(1 - \alpha_2)\}\,, \tag{5}$$

where $\alpha_1 = \Phi\left(z_{\alpha/2}(1 + m/n)^{1/2} + \widehat{z}_0(m/n)^{1/2}\right)$, $1-\alpha_2 = \Phi\left(z_{1-\alpha/2}(1 + m/n)^{1/2} + \widehat{z}_0(m/n)^{1/2}\right)$, Φ is the standard normal distribution function, and $\widehat{z}_0 = \widehat{\eta}_B(\widehat{K}_B(T_n))$. Similar to the BC confidence interval, (5) is derived under the assumption that there is a transformation g such that $g(T_n) - g(\theta_n) + z_0$ has a standard normal distribution. This is a reasonable assumption for $T_n = p$-value due to results like (1). However, consistency of these intervals should hold under weak consistency assumptions for the bootstrap distribution similar to those found in Theorem 4.1 of Shao and Tu (1996).

In practice, I suggest using $B = 999$ as in Examples 3 and 4, but $B = 99$ was used for some of the simulations in Section 4 (where only $B = 999$ results are reported), and the intervals performed reasonably well. For the binomial example, I calculated the endpoints of the intervals directly from the binomial distribution, i.e., from the exact bootstrap distribution. Note that in this paper I am using nonparametric bootstrap methods, resampling directly from the observed data. But for binomial data that can be strung out as a sequence of Bernoulli random variables, directly resampling of the Bernoulli data is equivalent to resampling from a binomial(n,$\widehat{\pi}$).

Where does the factor $(1 + m/n)^{1/2}$ come from? The easiest way to motivate it is to consider the following pivotal statistic used to get a prediction interval for the sample mean from a future sample, $T_m = \overline{X}$, with known variance σ^2,

$$\frac{\overline{Y} - \overline{X}}{\sigma \left(\frac{1}{n} + \frac{1}{m} \right)^{1/2}}.$$

Inverting this statistic gives the prediction interval

$$\left(\overline{Y} - \sigma m^{1/2}(1 + m/n)^{1/2} z_{1-\alpha/2}, \overline{Y} + \sigma m^{1/2}(1 + m/n)^{1/2} z_{1-\alpha/2} \right),$$

and the factor $(1 + m/n)^{1/2}$ appears naturally due to taking into account the variability of both $\overline{Y}$ and $\overline{X}$.

4 Monte Carlo Results

In this section I give results on empirical coverage and average length of the BC prediction intervals described in Section 3 for situations related to Examples 1 and 4 of Section 2 (Tables 4 and 5) and some results on the bootstrap and jackknife standard errors related to Examples 1,2, and 4 (Table 6). In each situation of Tables 4 and 5, 1000 Monte Carlo "training" samples are generated as well as a corresponding independent "test" sample. For each training sample, the intervals are computed and assessed as to whether they contain the p-value of the corresponding test sample. The number of bootstrap replications is $B = 999$

for all cases displayed in Table 5, but $B = 99$ was also used and gave similar results for most cases. All computations were carried out in R.

Table 4: Empirical coverages and lengths of bootstrap prediction intervals for p-values from the binomial test of $H_0 : \pi = 1/3$ versus $H_a : \pi > 1/3$.

	$\pi = 1/2$				$\pi = 2/3$			
	Lower 90%	Upper 90%	Interval 80%	Average Length	Lower 90%	Upper 90%	Interval 80%	Average Length
$n = 20$	0.92	0.93	0.85	3.00	0.93	0.94	0.87	4.47
$n = 50$	0.91	0.91	0.81	4.04	0.92	0.92	0.84	7.11
	75%	75%	50%	Length	75%	75%	50%	Length
$n = 20$	0.78	0.79	0.57	1.49	0.78	0.82	0.60	2.42
$n = 50$	0.75	0.76	0.51	2.00	0.78	0.78	0.56	3.65

Monte Carlo replication size is 1000. Bootstrap estimates are based on binomial$(n, \widehat{\pi})$. Standard errors for coverage entries are bounded by 0.016. Average length is on the base 10 logarithm scale; standard errors are a fraction of the the last decimal given. "Lower" and "Upper" refer to estimated coverage probabilities for intervals of the form (L,1] and [0,U), respectively.

Table 4 is for the binomial sampling of Example 1. Two alternatives are covered, $\pi = 1/2$ and $\pi = 2/3$ with $H_0 : \pi = 1/3$ as in the example. Here we see excellent coverage properties for even $n = 20$. Because of the discreteness, the endpoints of the intervals were purposely constructed from the binomial$(n, \widehat{\pi})$ (equivalent to $B = \infty$ resamples) to contain at least probability $1 - \alpha$. This apparently translated into slightly higher than nominal coverage. The average interval lengths are on the base 10 logarithm scale because interval lengths on the two sides of the p-value are not comparable (e.g., in Table 1 for $n = 50$ compare $0.011 - 0.000031$ to $0.29 - 0.011$).

Table 5 is for the Anderson-Darling goodness-of-fit test for normality illustrated in Example 4. The rectangles data in that example are well fit by an extreme value distribution. So the left half of Table 6 gives coverage probabilities for the p-value from the Anderson-Darling test for normality when the data are from an extreme value distibution. The right-hand-side of Table 5 is for data from a standard exponential distribution. For the extreme value

distribution, the coverages are not very good for small sample sizes but are reasonable by $n = 100$. For the exponential distribution, the coverages are not too bad for even $n = 20$, but the improvement is very minor for larger sample sizes up to $n = 100$. The convergence to normality of the Anderson-Darling statistic under an alternative is very slow and is likely driving the the slow convergence of the coverage of the prediction intervals.

Table 5: Empirical coverages and lengths of bootstrap prediction intervals for p-values from the Anderson-Darling test for normality

	Extreme Value				Exponential			
	Lower 90%	Upper 90%	Interval 80%	Average Length	Lower 90%	Upper 90%	Interval 80%	Average Length
$n = 20$	0.79	0.79	0.58	2.20	0.85	0.87	0.73	4.05
$n = 50$	0.79	0.87	0.66	3.91	0.86	0.89	0.75	7.53
$n = 100$	0.86	0.88	0.74	6.06	0.86	0.89	0.74	11.1
	75%	75%	50%	Length	75%	75%	50%	Length
$n = 20$	0.65	0.67	0.32	1.17	0.69	0.82	0.42	2.21
$n = 50$	0.66	0.73	0.39	2.00	0.73	0.74	0.46	4.05
$n = 100$	0.71	0.76	0.47	3.20	0.72	0.73	0.45	5.92

Monte Carlo replication size is 1000. Bootstrap replication size is $B = 999$. Standard errors for coverage entries are bounded by 0.016. Average length is on the base 10 logarithm scale, standard errors are a fraction of the the last decimal given. "Lower" and "Upper" refer to estimated coverage probabilities for intervals of the form (L,1] and [0,U), respectively.

Table 6 briefly assesses the bias and variation of the bootstrap and jackknife standard errors for $-\log_{10}(p\text{-value})$ for $n = 20$ for one alternative for the tests from Examples 1, 2, and 4. Both bootstrap and jackknife are relatively unbiased as evidenced by the ratios $\widehat{\text{SD}}/S$ near one, where $\widehat{\text{SD}}$ is the Monte Carlo average of the bootstrap and jackknife standard errors for $-\log_{10}(p\text{-value})$, and S is the Monte Carlo sample standard deviation of the $-\log_{10}(p\text{-value})$ values. The bootstrap, however, appears to have an advantage over the jackknife in terms of variability because the coefficient of variation (CV) of the standard errors is generally lower than that of the jackknife. This coefficient of variation is just the

Monte Carlo sample standard deviation of the standard errors divided by the Monte Carlo average of the standard errors. Other alternatives gave similar results with larger sample sizes showing improved results. For example, for the Anderson-Darling test with extreme value data at $n = 50$, the ratios $\widehat{\text{SD}}/S$ are 0.98 and 0.95, respectively, and the coefficients of variation are 0.42 and 0.57.

Table 6: Bias and coefficient of variation (CV) of bootstrap and jackknife standard errors of $-\log_{10}(p\text{-value})$ for the binomial test, t-Test, and Anderson-Darling goodness-of-fit test for normality when $n = 20$. Bias assessed by the ratio of $\widehat{\text{SD}}$=Monte Carlo average of the bootstrap and jackknife standard errors for $-\log_{10}(p\text{-value})$ to S=Monte Carlo sample standard deviation of $-\log_{10}(p\text{-value})$.

	Binomial		t-Test		Anderson-Darling	
Null	$H_0 : \pi = 1/3$		$H_0 : \mu = 0$		Normal	
Alternative	$H_a : \pi = 1/2$		$H_a : \mu = 1/2$		Extreme Value	
	$\widehat{\text{SD}}/S$	CV	$\widehat{\text{SD}}/S$	CV	$\widehat{\text{SD}}/S$	CV
Bootstrap	1.00	0.36	1.06	0.25	0.94	0.36
Jackknife	0.92	0.42	1.02	0.28	0.90	0.70

Monte Carlo replication size is 1000. Bootstrap replication size is $B = 999$.
Standard errors for all entries are bounded by 0.04.

5 Discussion

Frequentists typically accept p-values as a useful measure of evidence against a null hypothesis. But it is likely that most are not sufficiently aware of the inherent variability in p-values except at the null hypothesis where p-values are uniformly distributed for continuous statistics. The suggested prediction intervals are sobering in that regard, generally attenuating excitement about p-values in the range 0.005 to 0.05. More comforting, perhaps, the intervals tell us that in a replication of an experiment, *on average over all experiments* (not

over all replications for a given experiment), about half of the p-values will be below the observed p-value and half above because the intervals are roughly centered at the observed p-value. The prediction intervals also caution us about over-selling exact p-values in place of approximate p-values.

The size of the standard errors of $-\log_{10}(p\text{-value})$ is perhaps even more revealing. In all the examples I have considered (not just those reported here), when the observed p-value is below 0.05, these standard errors are of the same magnitude as the $-\log_{10}(p\text{-value})$, further supporting use of the star (*) method of reporting test results.

ACKNOWLEDGEMENT

I thank Len Stefanski for helpful discussions and comments on the manuscript and for the connection to the star (*) method.

REFERENCES

Lambert, D., and Hall, W. J. (1982), "Asymptotic Lognormality of P-Values," *The Annals of Statistics*, **10**, 44-64.

Larsen, R. J., and Marx, M. L. (2001), *An Introduction to Mathematical Statistics and Its Applications*, 3rd. ed., Prentice-Hall, Upper Saddle River: New Jersey.

Mojirsheibani, M., and Tibshirani, R. (1996), "Some Results on Bootstrap Prediction Intervals," *The Canadian Journal of Statistics*, **24**, 549-568.

Mojirsheibani, M. (1998), "Iterated Bootstrap Prediction Intervals," *Statistica Sinica*, **8**, 489-504.

Shao, J, and Tu, D. (1996), *The Jackknife and Bootstrap*, Springer: New York.

Stephens, M. A. (1986), "Tests Based on EDF Statistics" in D'Agostino, R. B. (ed.) and Stephens, M. A., *Goodness-of-fit Techniques*, Marcel Dekker Inc (New York), 97-193.