

5. ANSWERS TO PROBLEMS

1. Random sample of $n=1000$ from the normal distribution, mean 12, standard deviation 4.

- a) Each sample will be different but should be symmetric and have few “outliers” in the outlier box plot. Note the relatively narrow interquartile range.
- b) The sample mean and standard deviation should near 12 and 4, respectively. Skewness and kurtosis coefficients should be near 0.
- c) Each sample will be different but the fit will usually closely approximate a straight line.
- d) H_0 : The variable has a normal distribution in the population
 H_a : The variable does not have a normal distribution in the population
 The probability of rejecting H_0 is α , which is usually set to 0.05 by convention. Therefore about 1 student in 20 will reject the null hypothesis, but 19 of 20 will not.

2. Probabilities under the normal curve

- a) 0.159
- b) 0.025
- c) Since the normal curve is symmetric, $P(Z = 2.50) = P(Z = -2.50) = 0.006$.
 $P(Z = 2.50) = P(Z > 2.50)$ because the normal curve is a continuous distribution, and the probability of any single point is zero since a point has no width.
- d) $P(Z > -0.65) = P(Z < 0.65) = 0.742$
- e) $P(-2.3 < Z < 0.7) = 1 - 0.011 - 0.242 = 0.747$.
- f) $P(Z < -1.2 \text{ or } Z > 0.2) = 0.115 + 0.421 = 0.536$.
- g) Z can't be both less than -1.2 and greater than 0.2 so this probability = 0.
- h) -1.645
- i) 2.326
- j) -1.15 and 1.15

3. Fruitfly data, ignoring treatment groups

- a) The fit of male lifespan to the normal distribution appears reasonably good. The data have a distribution that appears flatter in the middle than the normal distribution (called negative kurtosis, or platykurtosis). The normal quantile plot appears linear.

b) Ho: Lifespan has a normal distribution in the population
 Ha: Lifespan does not have a normal distribution in the population
 $W = 0.975$, $P = 0.232$. Since $P > 0.05$ we fail to reject Ho. Rejecting Ho does not mean it is correct.

c) Male thorax appears to show a greater departure from the normal distribution than male lifespan. The distribution is not symmetric but rather is skewed to the left. The normal quantile plot appears non-linear.

Ho: Thorax has a normal distribution in the population
 Ha: Thorax does not have a normal distribution in the population
 $W = 0.912$, $P < 0.001$. Since $P < 0.05$, we reject Ho.

The normal and smooth curves seem to closely follow the data, but the histogram does seem to show a minor positive skew. The visual appraisal and statistical test seem to both indicate possible normality.

4. Fruitfly data, separate analyses for different treatment groups

- a) The fit of male lifespan to the normal distribution appears reasonably good for each treatment, although sample size is smaller. The normal quantile plots are roughly linear. Separate Shapiro-Wilks tests for each treatment group reveals that non depart significantly from the normal distribution (all $P > 0.05$).
- b) Males provided with 8 virgin females daily appear to have suffered the shortest lifespans, followed by males presented with 1 virgin female daily. If statistically significant (methods to test this will be covered later in the course) the pattern implies that mating is deadly for male flies.
- c) The box plots reveal asymmetry in the distribution of lifespans within some treatment groups (although recall that the null hypothesis of normality was not rejected). The interquartile ranges also appear broader than seen in random samples from a normal distribution (see (1a)).
- d) The normal quantile plots are all produced in a single figure, but in different colors. This can be somewhat confusing if data from different treatment groups overlap, but otherwise the overall picture is useful.

5. Distribution of sample means

- a) The data themselves (column “ $n=1$ ”) are highly skewed. Fit of sample means to the normal distribution improves with increasing sample size. The fit is still poor for $n=5$. The fit for $n=10$ is already much better. By $n=50$ the distribution of sample means is nearly normal. Visually, it would be difficult to detect a difference between the distributions of sample means for $n=50$ and $n=100$ and the true normal distribution using histograms, box plots, and normal quantile plots.

- b) The biggest changes to the distribution of sample means occur in skew rather than kurtosis. The data are highly positively skewed (long tail to the right), and the distribution of sample means for samples of size $n=5$ and $n=10$ still show positive skew. Skew is much reduced in means of samples of size $n=50$ and $n=100$. Kurtosis shows little pattern.
- c) For my samples (yours may be different), the first three variables were significantly nonnormal at an alpha level of 0.05. The distribution of my sample means at $n=50$ was not significantly different from the normal distribution, but the *P-value* was only 0.066. The distribution of sample means based on samples of $n=100$ failed to reject the null hypothesis of normality.
- d) The principle illustrated here is the Central Limit Theorem. This theorem states that the sum (or average) of a large sequence of random variables (e.g., the sample mean of a random sample of large n) will have an approximately normal distribution regardless of the distribution of the variable itself in the population.