Probability Plots

Keegan Korthauer
Department of Statistics
UW Madison

Recap

Summary Statistics

- Measuring central tendency
  - Mean
  - Median
  - Mode
- Measuring spread
  - Standard deviation
  - Percentiles
- Graphical summary
  - Histogram
  - Box plot

Probability

- Conditional probability
  - Total probability
  - Bayes rule
- Distributions
  - Discrete: Binomial, Poisson, Geometric
  - Continuous: Normal, Uniform, Exponential

Statistical Inference

- Point estimation
- Central Limit Theorem
- Confidence intervals
- Hypothesis testing
- Simple Linear Regression
- Multiple Regression

PROBABILITY PLOTS

How to construct

Interpretation

How to Choose a Distribution?

- So far we've considered two scenarios:
  1. We know what distribution our data follow and are given the parameter values
  2. We know that our data come from a certain distribution but do not know the parameter values so we estimate them (e.g. $\hat{p} = X / n$ in the binomial case) and their uncertainty
- A third scenario to consider: we *suspect* that our data follow a certain distribution, but we don't know for sure. **How do we check?**

Comparing Sample Distributions to Population Distributions

- Say we have a sample of 5 measurements that we suspect come from a normal distribution:

  3.01, 3.35, 4.79, 5.96, 7.89

- Compare the distribution of our sample with the distribution of the suspected population (normal, in this case) to see if they are similar using a **probability plot**!

Intuition Behind the Probability Plot

- For each original observation X_i in our sample, find its percentile
- For each of these percentiles, find the quantile Q_i of the suspected distribution that corresponds to it
- Examine the ordered pairs (X_i, Q_i)
  - If the data do come from the suspected distribution, they will lie close to a straight line
  - If they come from some other distribution, the points could be far from a straight line

How to Construct a Probability Plot

- First find the sample 'percentiles'
- For each of these find the quantile of the suspected normal distribution
  - for i=1, we find Q_1 such that $P(X \leq Q_1) = 0.1$ when $X \sim N(\mu, \sigma^2)$
  - best guess for $\mu = 5$ and $\sigma = 2$ (sample mean and standard deviation)
  - Standardize: $P(Z \leq (Q_1 - 5)/2) = 0.1$
  - From table: $P(Z \leq -1.28) \approx 0.1$
  - So $Q_1 = 2*(-1.28)+5 = 2.44$

i	X_i	"Percentiles" $(i-0.5)/n$	Q_i
1	3.01	0.1	2.44
2	3.35	0.3	3.95
3	4.79	0.5	5.00
4	5.96	0.7	6.05
5	7.89	0.9	7.56

Not true percentiles, but evenly spaced from 0 to 1

Plot X_i vs Q_i

Visualizing Q_i

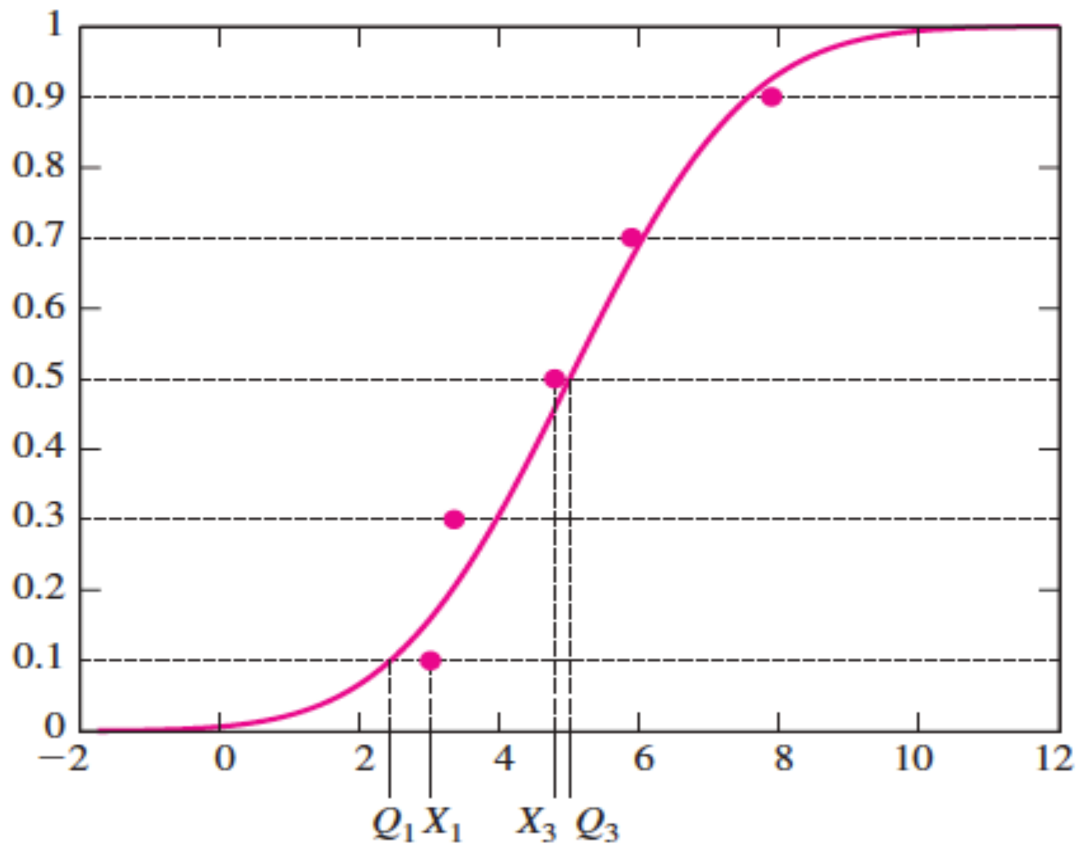


FIGURE 4.21 The curve is the cdf of $N(5, 2^2)$. If the sample points $X_1, \ldots, X_5$ came from this distribution, they are likely to lie close to the curve.

Probability Plots

The **probability plot** consists of the points (X_i, Q_i). Since the distribution that generated the Q_i was a normal distribution, this is called a **normal probability plot**.

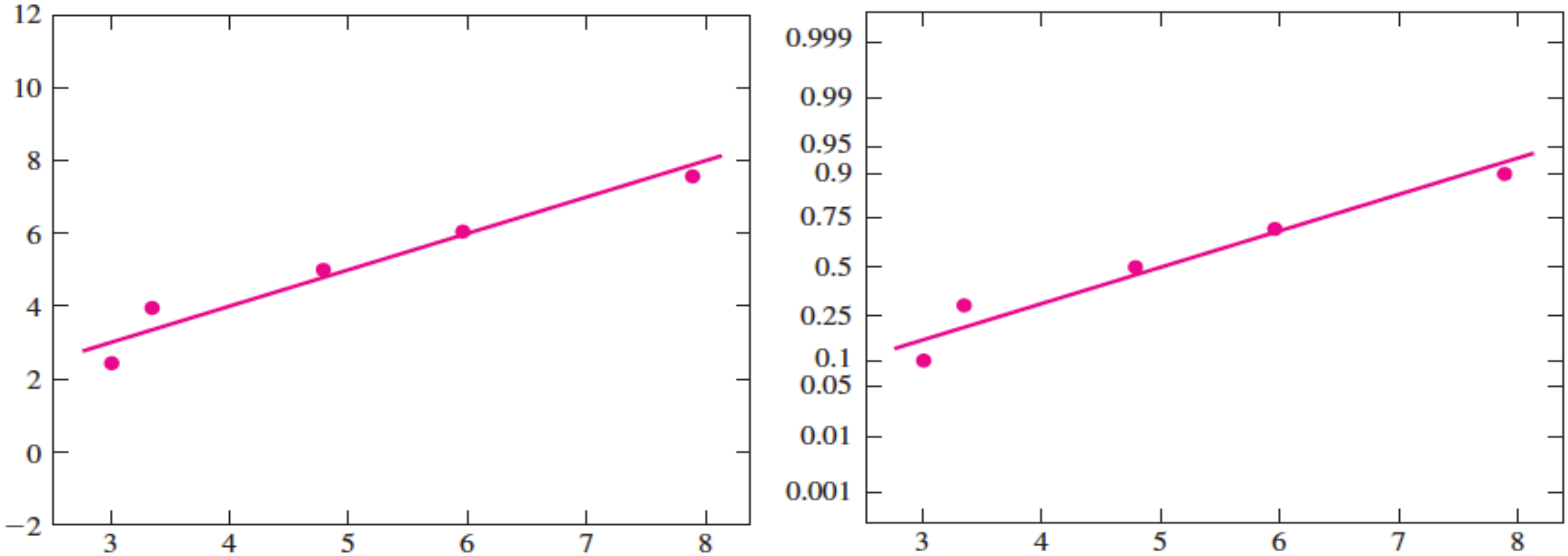
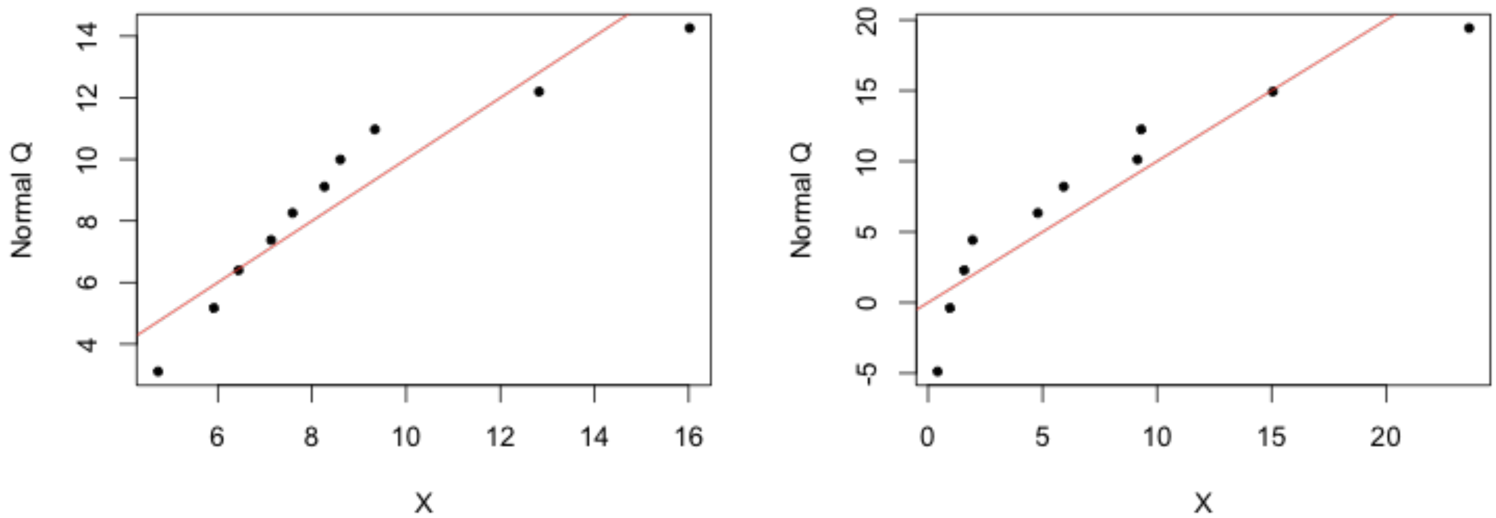


FIGURE 4.22 Normal probability plots for the sample $X_1, \ldots, X_5$. The plots are identical, except for the scaling on the vertical axis. The sample points lie approximately on a straight line, so it is plausible that they came from a normal population.

Sample Size

- With a small sample, probability plots will only show large departures from the suspected distribution
- Rule of thumb: sample size of **30 or more** will yield a reliable probability plot
- Use computer program (like R) to generate plots

Example – Sample of Size 10 vs 500



Example - Normal Probability Plots

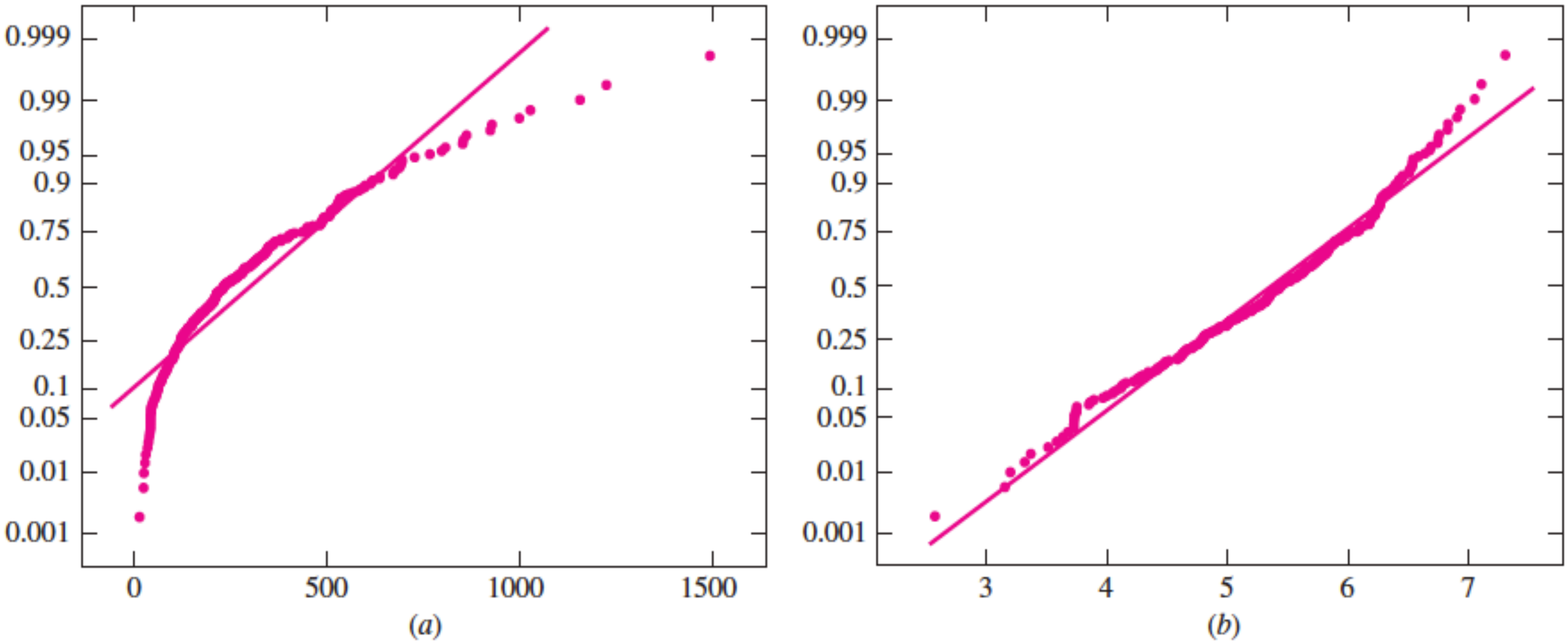
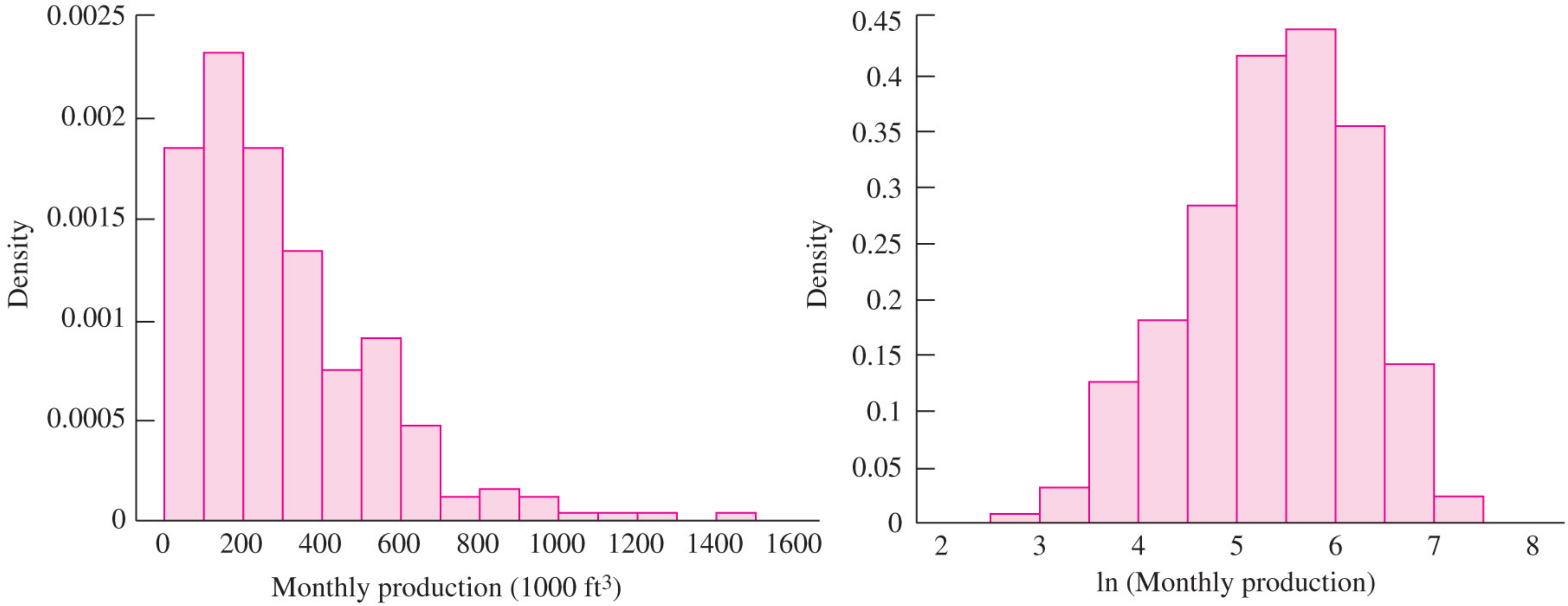


FIGURE 4.23 Two normal probability plots. *(a)* A plot of the monthly productions of 255 gas wells. These data do not lie close to a straight line, and thus do not come from a population that is close to normal. *(b)* A plot of the natural logs of the monthly productions. These data lie much closer to a straight line, although some departure from normality can be detected. See Figure 4.16 for histograms of these data.

Example - Normal Probability Plots



(a) (b)

Interpretation of Probability Plots

- No hard-and-fast rules – use the 'eye-ball' method
- Look for strong trends
- Common for a few points at the ends to stray
- Outliers will be far from the line when most of the others are close

Now What?

- Your plot shows strong departure from your suspected distribution. So what can you do?
  - Try plotting against the quantiles of a different distribution
  - Transform your data – more on this in Chapter 7
    - log-transformation
    - square root transformation
    - power transformation

Next

- Central Limit Theorem

- Introduction to R

- Exam 1 Review