Normal probability plots

- General idea: Order observations and plot against the standardized expected values of the observations assuming that the data are normally distributed.
- If data are normally distributed, observations will approximately equal their expected value under the normal distribution.
- Thus a linear (straight-line) trend on the "QQ-Normal" plot suggests that data are from an approx. normal distribution.

Graphics: Assessing departures from normality

Normal probability plots, "QQ-plots" (p. 79-83, M&M; p. 215-216 of *Sleuth*)

Idea behind a normal probability plot for data $X_1, \cdots, X_n$:

1. Arrange values from smallest to largest. Record what percentile of the data each value occupies. (ex: smallest observation in set of 20 is 5% point.)
2. Find z-scores at these percentiles. (z=-1.645 corresponds to 5% point of the standard normal.)
3. Plot x against z. If data is approximately normal, the result will be approximately a straight line with slope σ and intercept μ.
4. Features
   - Expect some scatter around the line even in the case of normality.
   - Identify heavy/light left/right tails
   - Identify outliers
   - Look for systematic departures from normality

A *t*-Ratio for Two-Sample Inference

- Data: $X_1, \cdots, X_n; Y_1, \cdots, Y_m$
- *Assumptions*:
  - X and Y populations are approximately normally distributed.
  - The two populations have approximately equal standard deviations, $\sigma_X = \sigma_Y = \sigma$.
  - Independence between and within groups of subjects.
- Two-sided test: $H_o : \mu_X = \mu_Y$ versus $H_A : \mu_X \neq \mu_Y$
- Test statistic: $T = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{SE(\bar{X}-\bar{Y})}$
- Estimate $SE(\bar{X}-\bar{Y})$ using:
  $SE(\bar{X}-\bar{Y}) = s_p\sqrt{\frac{1}{n}+\frac{1}{m}}$
  $s_p = \sqrt{\frac{(n-1)s_X^2+(m-1)s_Y^2}{n+m-2}}$, d.f. $= n+m-2$
  s_p:*pooled estimate* of σ^2, valid if $\sigma_X = \sigma_Y = \sigma$.

Two independent samples: Inference

- Two-sided test: Reject H_o at level α if $|T| > t_{1-\alpha/2,n+m-2}$
- One-sided test:
  For $H_A : \mu_X - \mu_Y < 0$, Reject H_o at level α if $T < -t_{1-\alpha,n+m-2}$
  For $H_A : \mu_X - \mu_Y > 0$, Reject H_o at level α if $T > t_{1-\alpha,n+m-2}$
- Confidence interval for $\mu_x - \mu_y$ at level $1-\alpha$ is:

$$(\bar{X}-\bar{Y}) \underline{+} t_{1-\alpha/2,n+m-2} SE(\bar{X}-\bar{Y})$$

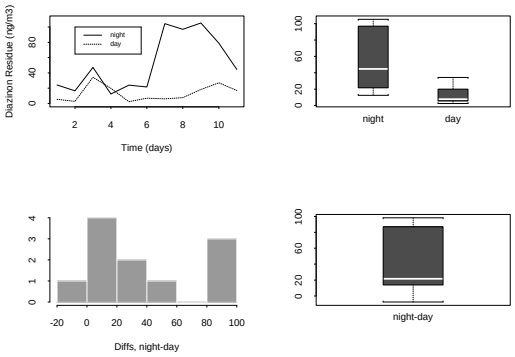
Matched Pairs: Example

Pesticides applied to an extensively grown crop can result in inadvertent area-wide air contamination. *Environmental Science and Technology* (October 1993) reported on air deposition of the insecticide diazinon on dormant orchards in the San Joachin Valley, CA. Ambient air samples were collected and analyzed at an orchard site for each of 11 days during the most intensive period of spraying. The levels of diazinon residue (in ng/m^3) during the day and at night were recorded.

- **How do mean levels of diazinon differ during the day and at night?**

obs	1.0	2.0	3.0	4.0	5.0	
day	5.4	2.7	34.2	19.9	2.4	
nite	24.3	16.5	47.2	12.4	24.0	
obs	6.0	7.0	8.0	9.0	10.0	11.0
day	7.0	6.1	7.7	18.4	27.1	16.9
nite	21.6	104.3	96.9	105.3	78.7	44.6

Matched Pairs: Diazinon Data



- Do diazinon levels increase at night?
- Give a 95% confidence interval for the difference in mean diazinon residue between night and day.
- What assumptions are necessary? Are these realistic?

Graphics: Comparison

- Features of a distribution: shape (normal? symmetric? skewed? concentrations or gaps in data? bimodal or unimodal?); center; spread (long-tailed?, outliers?). Also: missing data? censoring? truncation?
- Boxplot
  - center, spread, skew, outliers
  - comparison of several samples
- Histogram
  - relative frequencies of different values
  - mode of the data
  - spread, skew, peakedness
  - choice of bin size (see Week 1 applet)

Summary Statistics

- Summary statistics for NIGHT - DAY:

```
1st Qu. Median  Mean 3rd Qu.
  14.2   21.6  38.91   69.25
```

- s=36.58
- Resistant statistics
- How to changes in location or scale affect each of these statistics?
- Coefficient of variation

Matched Pairs: Rationale

- Increased precision in assessing differences by confining comparisons within matched pairs; variation lower within each pair than between the 2 groups
- Reduces the effect of confounding variables by screening out extraneous effects
- An example of *blocking*
  - *Block:* a group of experimental units or subjects that are known before the experiment to be similar in some way that is expected to affect the response to the treatments
  - Examples: twin studies, before/after,
  - Block design: random assignment of units to treatments separately within each block
- Consider the distribution of differences *within each block*

Matched Pairs: Inference

- Data: $(X_1, Y_1), \cdots, (X_n, Y_n)$ measurements on n pairs.
- Null hypothesis: $H_o : \mu_D = \mu_1 - \mu_2 = 0$
- Alternative: $H_A : \mu_D \neq 0$ (two sided); $H_A : \mu_D > 0$ or $H_A : \mu_D < 0$ (one sided)
- Form sample differences: $D_1 = Y_1 - X_1, D_2 = Y_2 - X_2, \cdots, D_n = Y_n - X_n$
- Estimate σ_D using:

$$s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2, \text{ where } \bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$$

- Test statistic: $T = \frac{\bar{D}_n - \mu_D}{s_D/\sqrt{n}}$
- Under H_o, $T \sim t_{n-1}$
- What assumptions are needed for the underlying population of differences?