

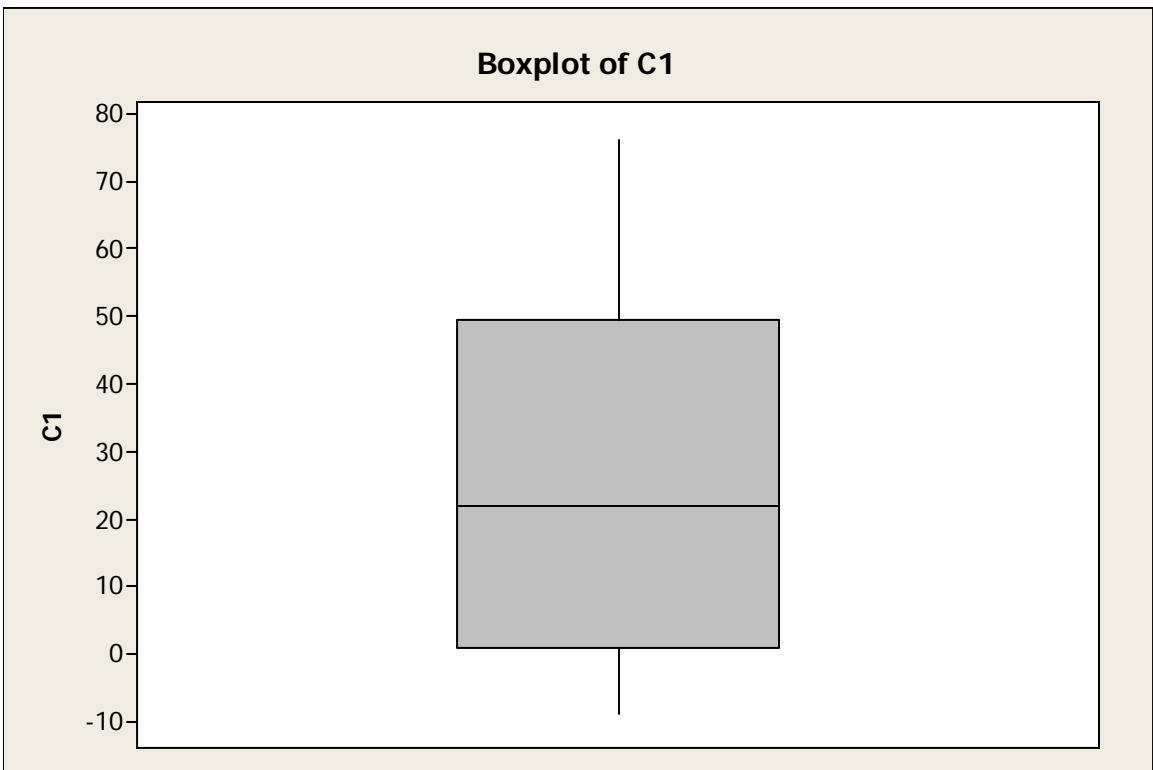
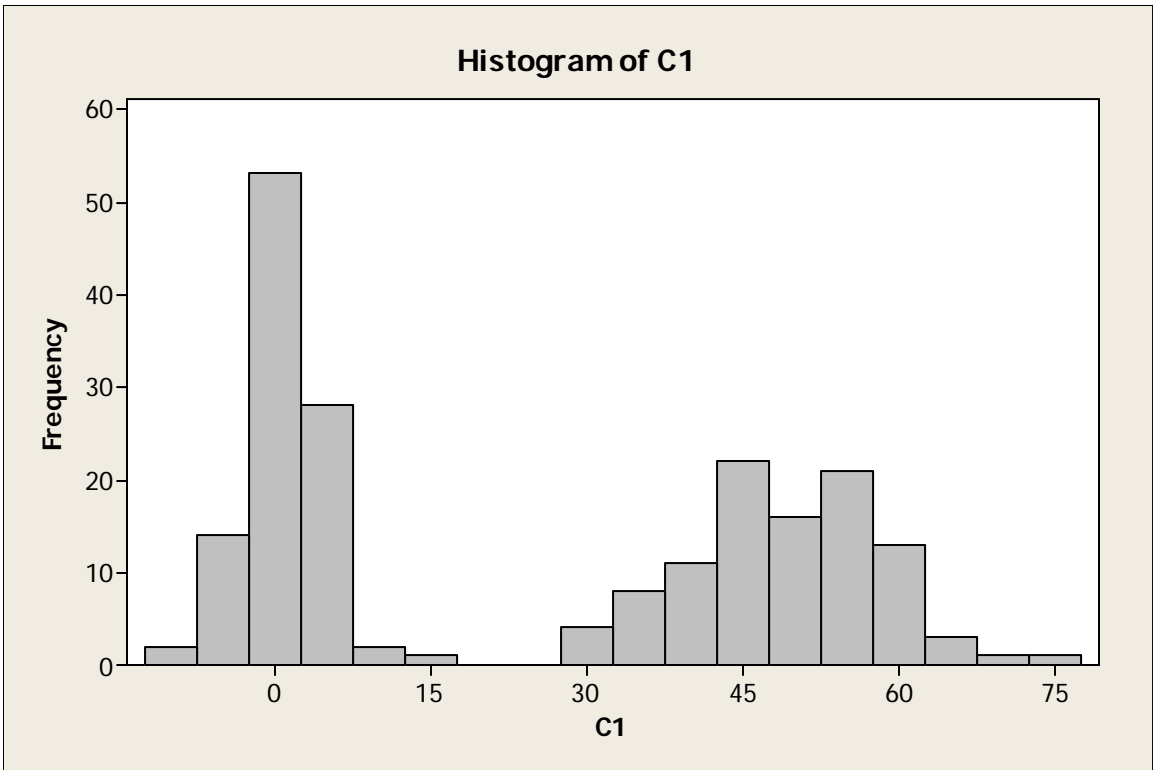
Thinking Statistically
Sat Gupta
Professor of Statistics
UNCG

Types of Studies

Descriptive Studies:

Just organize the data properly and present it nicely. The objective is not to draw any inferences. You should comment on important features of the data. It is easy to miss some of the important features!

The graphs below are based on merging two very distinct data sets. You will miss this aspect if you rely on one graph (the box plot) but will catch it if you use a histogram.



Observational Studies:

Analyze the available data and indicate trends. Can't make inferential statements.

Experimental/Inferential Studies:

Set up the experiment carefully, analyze the data and make inferential statements.

Mathematical Thinking vs. Statistical Thinking

Suppose the average GPA of male students at a college is 3.10 and the average GPA for female students is 2.90. These numbers are based on large samples. Consider the following two questions:

Is 3.10 greater than 2.90? (This is a Mathematical Question)

Is 3.10 significantly greater than 2.90? (This is a Statistical Question)

Statistical Significance

A finding is statistically significant if the probability of it happening just by chance is very small; say less than 5% or less than 1%.

Roughly speaking, this probability is called “p-value”. So, a small p-value typically indicates departure from status quo.

A result may be statistically significant but may not be practically significant.

Probability of three Heads in row when tossing a fair coin repeatedly is $(.5)(.5)(.5) = .125$. This is not a significant occurrence since it can happen 12.5% of the times, but probability of 5 heads in a row is $.03125$. So, 5 heads in a row happen only about 3% of the times, and hence this event is a statistically significant event. Neither event may be practically significant is the coin toss is only a simple game and carries no award.

Data Types

Numerical Data:

Height, Weight, Time, Number of Decayed Teeth.

Categorical Data:

Honey Bee Type (Queen/worker), Male/Female.

Ages of employees who were laid off

44, 45, 53, 61, and 62

Ages of employees who were not laid off

58 40 47 36 41 46 52 44 44 39 45 48 51 45 44
48 50 56 58 40 46 36 37 38 39 40 40 41 42 43
44 44 44 44 45 46 49 50 52 54 54 54 58 39 41
44 47 50 45 45 32 45 31 31 35 37 39 39 41 42
42 42 42 42 44 44 45 45 46 46 46 47 47 48 50
51 53 53 54 56 58 59 31 45 46 53 53 43 45 36
42 44 49 50 52 53 56 49 52 57 38 47 51 53 56
58 47 34 47 43 39 46 59 37 41 45 45 47 47 50
56 62 56 26 27 33 35 39 40 44 49 38 39 44 44
44 45 45 35 36 41 41 44 46 47 48 49 53 53 53
61 57 41 36 51 52 53 42 46 40 29 37 39 50 52
41 39 44 27 57 37 42 43 45 49 50 50 52

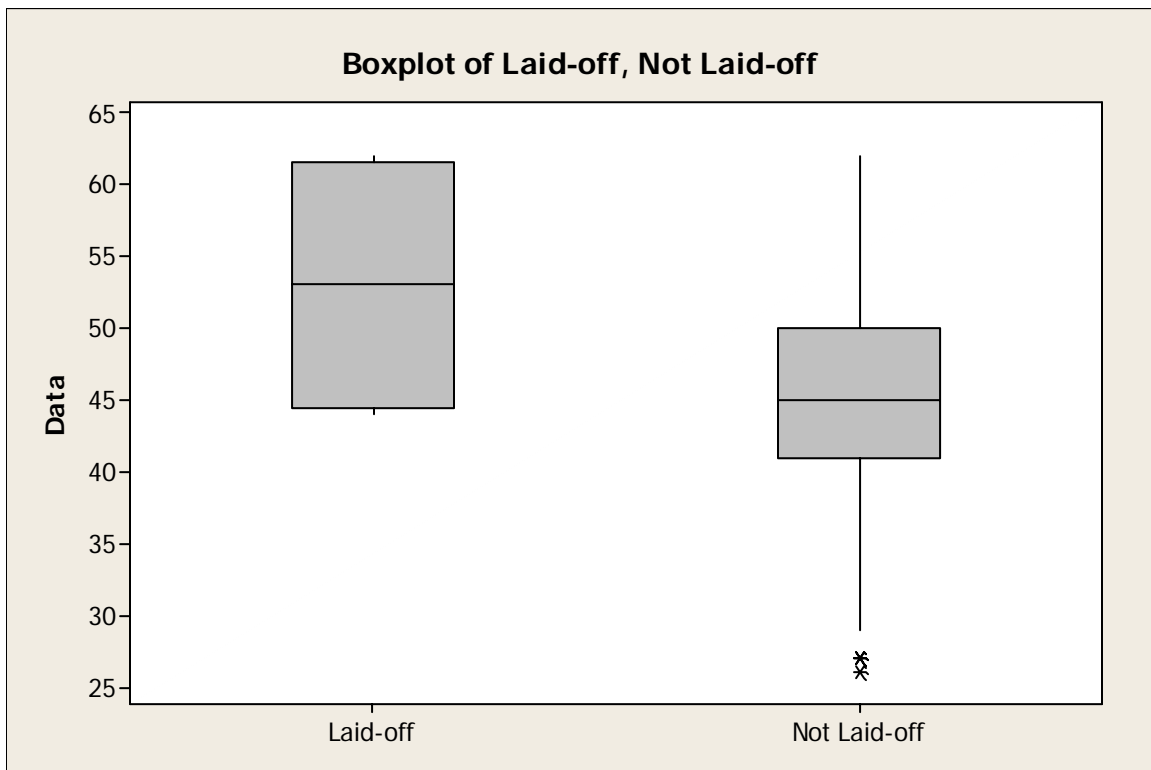
WAS THE LAY OFF PROCESS AGE NEUTRAL??

**Is it an observational study or an experimental study?
What are the consequences?**

Average Age of those who were laid off **53.00**

Average age of those who were not laid off **45.37**

A Visual Check



“Good data” or “not so good data”?

**Both groups show normality. The two outliers may not be very serious!
But this can be checked.**

Quantitative Analysis

Several options, depending on data characteristics: “good data” vs. “not so good data”, “numerical data vs. categorical data”.

Two-Sample Rank Sum Test (A Non-parametric Test):

Mann-Whitney Test and CI: Laid-off, Not Laid off

	N	Median
Laid off	5	53.00
Not Laid off	178	45.00

Two-sided p-value = .058

There is about 5.8% chance that the difference in the average ages of the two groups would be as wide as $53 - 45.37 = 7.63$ years, just because of normal statistical fluctuation. Since this value is more than .05, the age difference in the two groups should not be taken seriously.

A usual p-value threshold is .05.

Important Question: Are we testing whether the two means are similar, or are we testing if the mean for the laid off group is significantly higher than the not laid off group?

Should we have looked at a one-sided p-value?

Two Problems:

A non-parametric test is less likely to detect real group differences!

A two-sided test has larger p-value and hence is less likely to detect real group differences!

Data were good enough to allow parametric analysis. Also, a one-sided comparison is more appropriate here since we would be more interested in knowing if the older group was at a greater risk, and not if the risks were equal. The one-sided p-value for the Rank Sum Test is .029

**Two-sample t-test (A Parametric Test):
One-sided p-value = .01**

Chi Square Test Working with Count Data

Motivation: Federal Guidelines for age discrimination cases.

Using 60 as a cut off date, check for significance of association between age and lay off decision

		<u>Decision</u>	
		<u>Laid off</u>	<u>Not Laid off</u>
<u>Age</u>	<u>≥ 60</u>	2	2
	<u>< 60</u>	3	176

Odds Ratio (for lay off) 5.67

An employee in the “60 or older” age group was 5.67 times as likely to be fired as was an employee in the “under 60” age group. **Is 5.67 statistically significant?**

Fisher’s Exact Test p-value: .0035

Using 40 as a cut off age

		<u>Decision</u>	
		<u>Laid off</u>	<u>Not Laid off</u>
<u>Age</u>	<u>≥ 40</u>	5	143
	<u>< 40</u>	0	35

Fisher's Exact Test p-value .3416

Recall: Subgroup protection is also enforced.

Binary Logistic Regression

Try to establish a relation between ages and lay off status on a continuous scale

Let p_x be the lay off probability for an employee with age x . Can we predict p_x knowing x ?

Try setting up the simple linear regression model

$$p_x = \alpha + \beta x + \varepsilon_x$$

The problem is that left hand side has to be between 0 and 1 and right hand side is not restricted to be in this range!

Use the logit transformation on the dependent variable p_x .

$$\ln\left\{\frac{p_x}{1-p_x}\right\} = \alpha + \beta x + \varepsilon_x$$

Note that the range of $\ln\left\{\frac{p_x}{1-p_x}\right\}$ stretches from $-\infty$ to $+\infty$

There are some other such transformations available. We will discuss the issues involved in transformation selection in the next case study.

$$\hat{p}_x = \frac{\exp\{\hat{\alpha} + \hat{\beta}x\}}{1 + \exp\{\hat{\alpha} + \hat{\beta}x\}}$$

Age p-value= .028

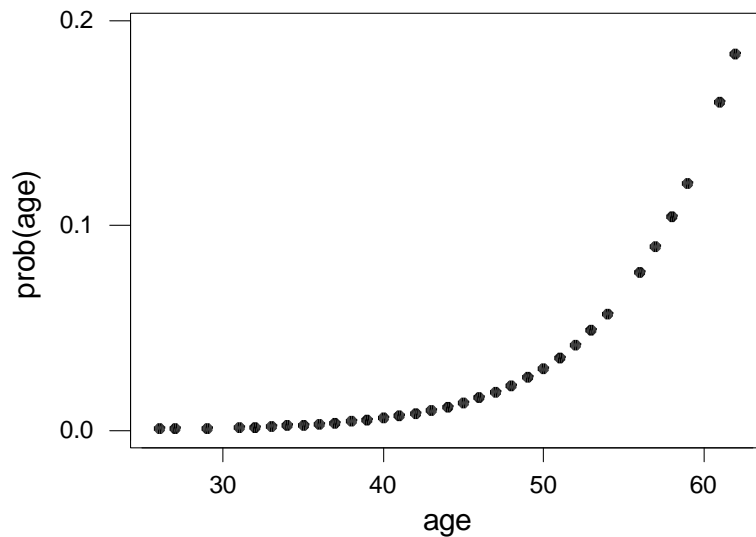
Age is a significant predictor of lay off decision.

Odds ratio = 1.18

Risk of lay off goes up 18% with each increment of one year in age.

Estimated lay off probability for various ages.

Figure 1: Probability of termination for different ages



The graph suggests that the probability of job termination increases exponentially with age, increasing from a low of .0006 for a 26 years old to .1836 for a 62 years old resulting in a $(.1836-.0006)/.0006 = 30,500$ % increase.

Important Issues:

- **‘Good data’ or ‘not so good data’**
Normality, no outliers, etc.
- **Study type**
Observational or Experimental
- **Appropriate hypotheses**
One-sided or two-sided etc.
- **Choice of data analysis method**
Parametric or non-parametric
- **Correct interpretation of results**
Indicates trend or causality

Milk Testing Protocol Set by FDA

Sensitivity = $P(T^+ | Disease)$

Specificity = $P(T^- | No Disease)$

The following response rates, all give same sensitivity estimates but with different confidence levels

10/10

20/20

30/30

100/100

Both sensitivity (at tolerance level) and specificity (at control level) of diagnostic test kits should be at least .90 with a confidence level of 95%.

In other words, lower 95% confidence limit in each case should be more than .90

If you follow this approach, you need to use One-sided (EXACT) Confidence Interval and not a two-sided CI.

Example:

Number of positive samples = 90

Number of positive test results = 85

Sample sensitivity = $85/90 = .944$

Using Normal Approximation

Rule of Thumb for approximation validity ($n \geq 30, n\hat{p} \geq 5, n(1 - \hat{p}) \geq 5$)

Two-sided 95 % lower confidence limit

$$\hat{p} - 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} = .944 - .069 = .875$$

PRODUCT FAILS

One-Sided 95% Lower Limit

$$\hat{p} - 1.645\sqrt{\frac{\hat{p}\hat{q}}{n}} = .944 - .04 = .904$$

PRODUCT PASSES

Exact One-sided 95% Lower Confidence Limit

Solve the following equation for p with $n=90, x=85$

$$\sum_{i=x}^n nCi p^i q^{n-i} = .05$$

Solution is $p=.887$

FAIL

CORRECT RESULT

Real Data

Dose/Response Methodology

Amoxicillin Tolerance Level 10 ppb

Concentration level (ppb)	0	3	4	6	8	10
# positive results	0/60	0/30	0/30	3/29	15/30	29/29

Ignoring all levels except 10 ppb

Exact 95 % Lower confidence limit = .902

PASS

Dose/Response Binary Regression

Logistic Model

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta \log x + \varepsilon_x$$

Probit Model

$$\phi^{-1}(p) = \alpha + \beta \log x + \varepsilon_x$$

Gompertz Model

$$\log[-\log(1-p)] = \alpha + \beta \log x + \varepsilon_x$$

FDA insisted on using Probit Model (ALWAYS)

**95 % lower confidence limit on sensitivity at 10 ppb
corresponding to each of the three models**

Logistic **.897** **Almost Pass!**

Probit **.873** **Fail (FDA)**

Gompertz **.926** **Pass**

**EXACT 95% Lower
Limit at 10 ppb
ignoring all other
dose levels** **.902** **PASS**

Goodness-of-fit Results

Logistic	Pearson Chi-square p-value	.1234	
Probit	Pearson Chi-square p-value	.0563	FDA
Gompertz	Pearson Chi-square p-value	.3692	

Probit transformation produces the worst quality of fit.

Pollen Removal Data
35 Bumblebee Queens, 12 Honeybee Workers

Variables:

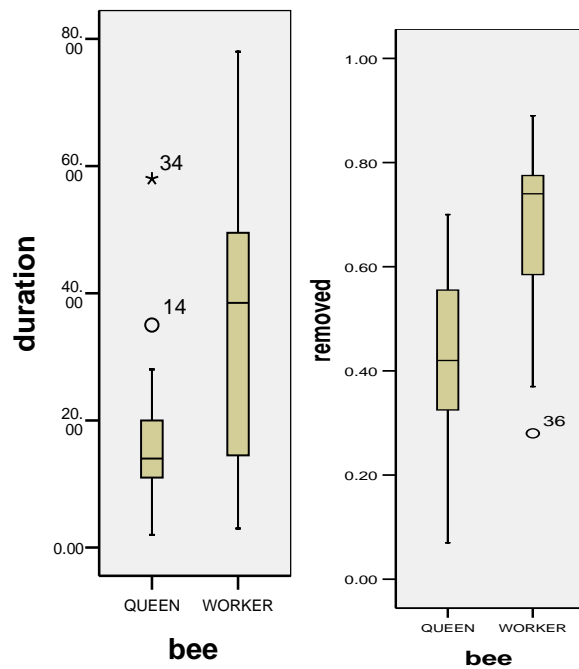
Proportion of pollen removed: Continuous on (0, 1)

**Time (in seconds) spent
on the flower: Continuous**

Bee Type (Queen or Worker): Categorical

<u>p</u>	<u>t</u>	<u>bee type</u>
.07	2.00	QUEEN
.10	5.00	QUEEN
.11	7.00	QUEEN
.12	11.00	QUEEN
.15	12.00	QUEEN
.19	11.00	QUEEN
.28	9.00	QUEEN
.31	9.00	QUEEN
.30	16.00	QUEEN
.34	17.00	QUEEN
.35	12.00	QUEEN
.39	14.00	QUEEN
.38	23.00	QUEEN
.40	35.00	QUEEN
.42	21.00	QUEEN
.40	10.00	QUEEN
.41	9.00	QUEEN
.42	7.00	QUEEN
.48	11.00	QUEEN
.48	13.00	QUEEN
.47	14.00	QUEEN
.49	16.00	QUEEN
.50	14.00	QUEEN
.51	17.00	QUEEN

.53	22.00	QUEEN
.58	13.00	QUEEN
.59	13.00	QUEEN
.65	12.00	QUEEN
.60	19.00	QUEEN
.60	23.00	QUEEN
.69	21.00	QUEEN
.70	27.00	QUEEN
.70	28.00	QUEEN
.51	58.00	QUEEN
.70	15.00	QUEEN
.28	3.00	WORKER
.37	12.00	WORKER
.52	10.00	WORKER
.65	17.00	WORKER
.76	24.00	WORKER
.89	33.00	WORKER
.74	44.00	WORKER
.70	46.00	WORKER
.79	48.00	WORKER
.78	51.00	WORKER
.74	64.00	WORKER
.77	78.00	WORKER



Report

bee		removed	duration
QUEEN	Mean	.4263	16.1714
	N	35	35
	Std. Deviation	.18197	10.01318
WORKER	Mean	.6658	35.8333
	N	12	12
	Std. Deviation	.18303	23.22942
Total	Mean	.4874	21.1915
	N	47	47
	Std. Deviation	.20888	16.68068

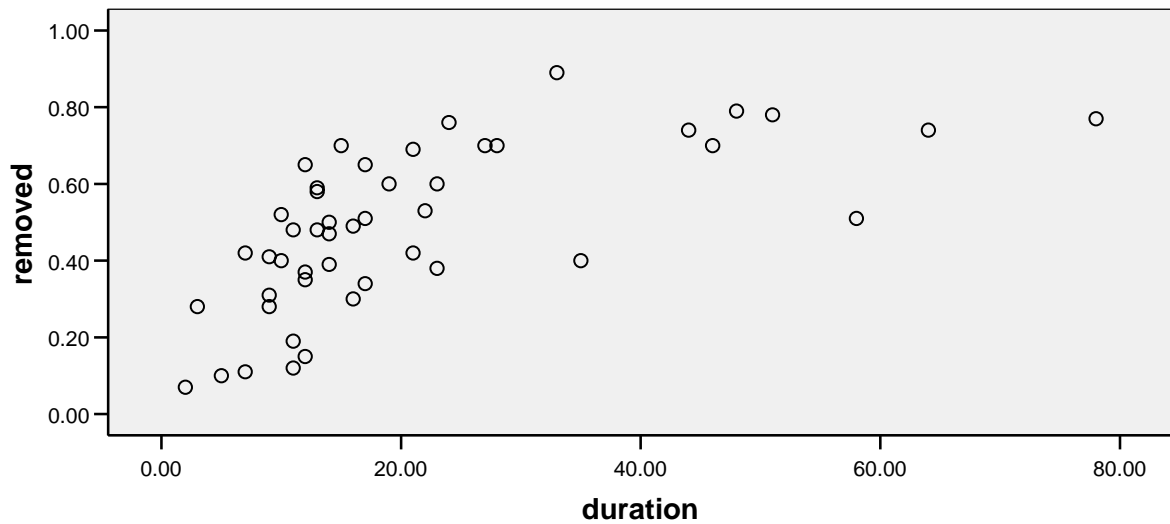
		Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
					Lower	Upper
removed	Equal variances assumed	.000	-.23955	.06096	-.36233	-.11677
	Equal variances not assumed	.001	-.23955	.06114	-.36750	-.11159
duration	Equal variances assumed	.000	-19.66190	4.82058	-29.37105	-9.95276
	Equal variances not assumed	.014	-19.66190	6.91606	-34.67311	-4.65070

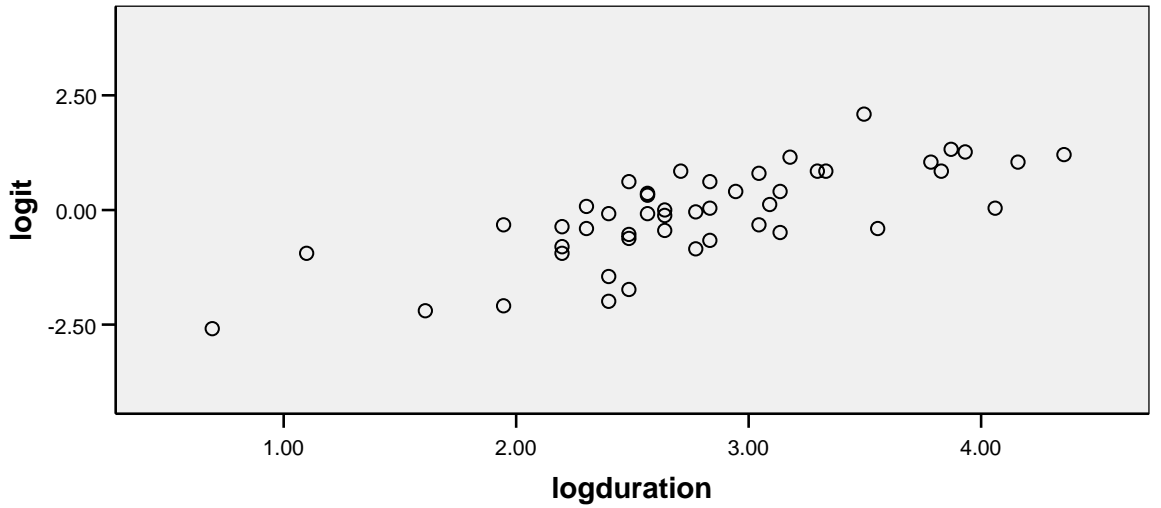
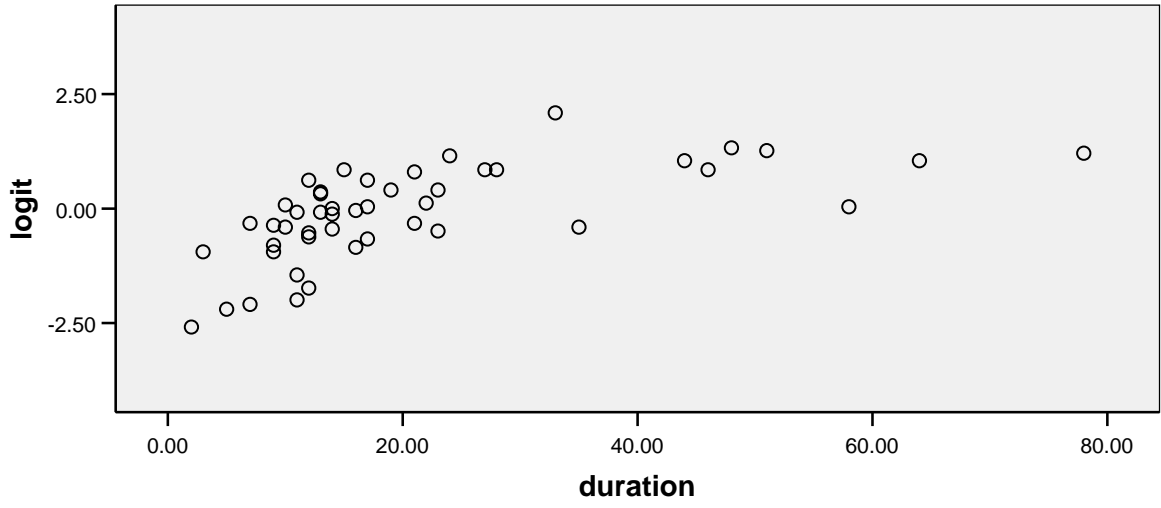
It is hard to tell if, on an average, the queens remove a smaller proportion of pollen as compared to workers! Time effect has to be factored in.

We could consider fitting a regression model of the type

$$\mu(\text{proportion} / \text{bee type}, \text{visit duration}) = \beta_0 + \beta_1(\text{bee type}) + \beta_2(\text{visit duration})$$

There will be some issues that need to be addressed. Some transformations may be needed. Once that is done, we can study the impact of bee type on proportion of pollen removed, **after time duration is neutralized.**





Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2.145	.496		-4.321	.000
	bee	-.570	.236	-.247	-2.409	.020
	logduration	.889	.140	.650	6.339	.000

a. Dependent Variable: logit

How do we interpret these results?

The fitted model is

$$\mu\left[\ln\left\{\frac{p}{1-p}\right\}\right] = -2.145 - .570(\text{bee type}) + .889(\text{log duration})$$

Note that for a normal distribution, mean = median

This means

$$\text{Mean} [\log (f)] = \text{Median} [\log (f)] = \log \{\text{median} (f)\}$$

From the equation above,

$$\text{Log}\{\text{Median} [p/(1-p) \text{ for queens}]\} - \text{Log}\{\text{Median} [p/(1-p) \text{ for workers}]\} = -.570$$

$$\text{Median} \{p/(1-p)\} \text{ for queens} = e^{-.570} \cdot \text{Median} \{p/(1-p)\} \text{ for workers} \\ (.565)$$